Carbon dioxide at extreme conditions : liquid(s), crystals, glasses and their transformation from ab initio topological methods

Mathieu Moog

Université Sorbonne Université

École Doctorale 397:
Physique et Chimie des Matériaux

Carbon dioxide at extreme conditions

Liquid(s), crystals, glasses and their transformations from *ab initio* topological methods

---

Thèse de Doctorat de Physique

presentée par

Mathieu Moog

Pour l'obtention du grade de:

Docteur de Sorbonne Université

Dirigée par

A. Marco Saitta & Fabio Pietrucci

Défendue le 20 Septembre 2019, devant un jury composé de :

Dominique COSTA	Rapportrice
Sandro SCANDOLO	Rapporteur
Marie-Laure BOCQUET	Examinatrice
Jean-Louis HAZEMANN	Examinateur
Edouard KIERLIK	Examinateur
A. Marco SAITTA	Directeur de Thèse
Fabio PIETRUCCI	Directeur de Thèse

August 29, 2019

Remerciements

Résumé

This thesis was written in english as one of its examinator is not francophone and in the spirit of accessibility to the larger number. This choice is possible in accordance with the internal rules of Doctoral School 397, where the author is registered, with the caveat *that a summary of around 10 page be written in french. Therefore, the following pages constitute said summary.*

Nous avons décidé d'écire cette thèse en anglais car l'un des rapporteurs de cette thèse n'est pas francophone, mais aussi par soucis d'accessibilité au plus grand nombre. Ce choix est rendu possible par le règlement de l'école doctorale 397, où l'auteur de cette thèse est inscrit, sous réserve qu'un résumé d'environs 10 pages soit écrit en français. C'est ce résumé qui est présenté dans les pages suivantes.

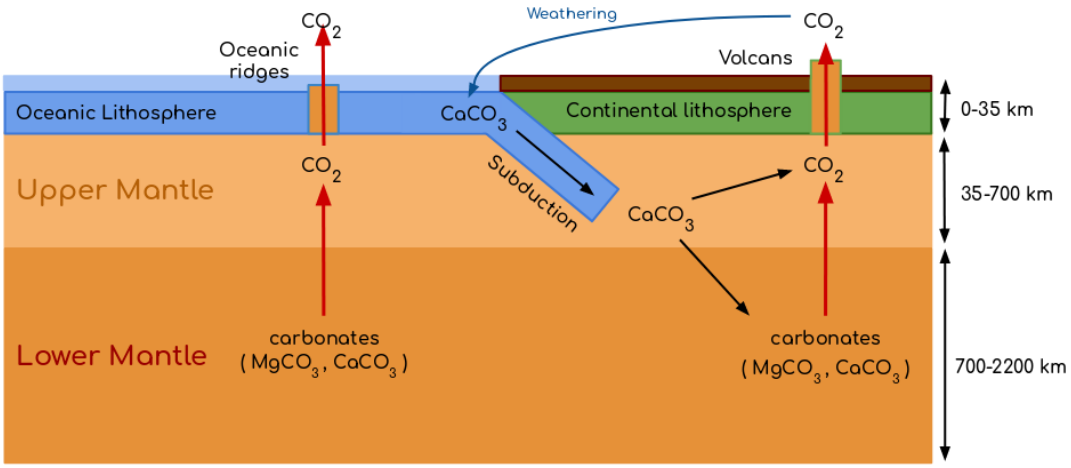


Figure 1: Cycle du carbone geothermique

Le dioxyde de carbone est une molécule dont les effets sur l'atmosphère comme gaz à effet de serre sont les mieux connus et les plus étudiés. Cependant, cette molécule joue aussi un rôle très important dans les phénomènes géologiques [1, 2], et notamment dans le cycle du carbone profond [3] (figure 1). En effet, il se forme dans le manteau supérieur de la Terre et est l'un des constituants majoritaires des émissions volcaniques, permettant la ré-introduction en surface du carbone enfouis par la subduction de couches géologiques contenant des éléments organiques (figure 1). Qui plus est, il a été récemment mis en évidence que le dioxyde de carbone pouvait se former jusque dans le manteau inférieur, à des prondeurs situées entre 750 et 1500km sous le niveau de la mer[4, 5]. Cette formation en profondeur pourrait avoir une influence pour la chimie du manteau inférieur de la Terre, comme par exemple sur la formation de diamants[4]. La compréhension du comportement du dioxyde de carbone dans les conditions géothermiques du manteau inférieur revêt donc

une importance particulière pour la compréhension de ce dernier.

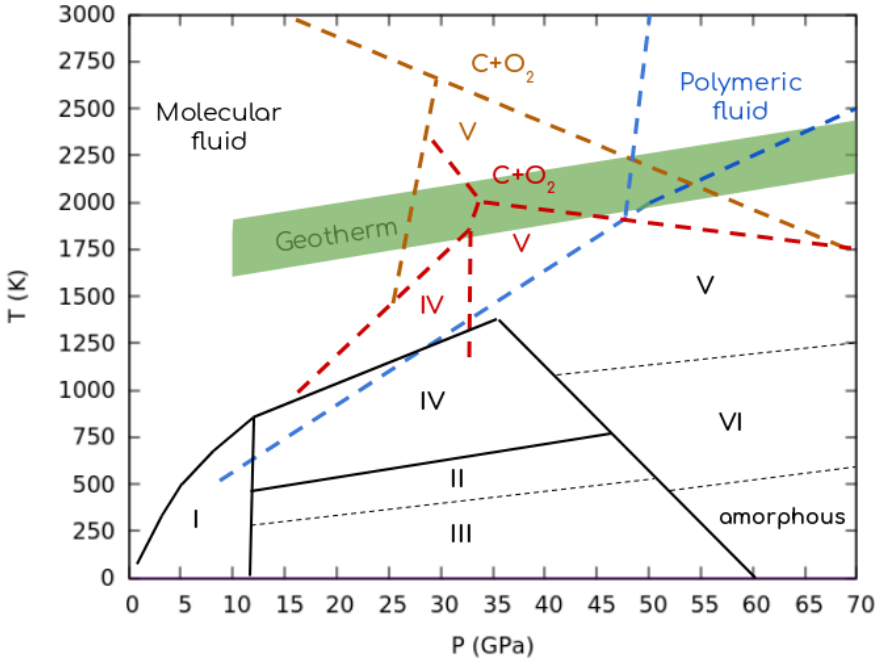


Figure 2: Diagramme de phase du CO_2 sous haute pression, établit en suivant [6]. En bleu la ligne de fonte des phases cristallines du CO_2 et ligne de transition liquide moléculaire - liquide polymérique, d'après [7]. En rouge les limites entre les phases IV,V fluid moléculaire et dissociation du CO_2 d'après [8]. En marron les limites entre la phase V, le fluid moléculaire et la dissociation du CO_2 d'après [9]. Les lignes pointillées indiquent des limites cinétiques entre les phases.

En plus de ces aspects, le dioxyde de carbone a été un des premiers systèmes à être étudié sous haute pression [10, 11, 12]. En effet, il s'agit d'une molécule simple et abondante dans la nature, et qu'il est assez facile de crystalliser en glace carbonique. Le dioxyde de carbone a donc été un objet d'étude de choix dans ce domaine de recherche [13, 14, 15]. Les nombreux travaux ont mis à jour un diagramme de phase haute pression très riche, contenant cinq phases moléculaires cristallines (I [11, 16], II [13, 17, 18], III [14, 16, 19], IV [20, 21, 22, 23], VII [24]), deux phases polymériques cristallines (V [25, 26, 27, 28, 29, 30, 31, 32] et VI [27, 33, 34, 35, 36]) ainsi que trois phases désordonnées: un liquide moléculaire [6, 24], un liquide polymérique [7, 37] et une phase amorphe polymérique [38, 39, 40] (figure 2).

Si les structures des différentes phases ont longtemps été l'objet de controverses [23, 33, 41, 42], le rapprochement d'approches experimentales et théoriques [6, 38, 43, 44] a permit d'établir suffisament de consensus pour qu'il soit possible de décrire le diagramme de phase du dioxyde de carbone sous haute pression, au moins dans les grandes lignes (figure 2). Ce diagramme peut grossièrement être divisé en deux zones en fonction du type de structure qu'on y trouve. Dans la région de pression comprise entre 0 et 40 GPa

(figure 2) le CO_2 est purement moléculaire. Au-dessus de 40 GPa, le dioxyde de carbone forme uniquement des phases polymériques, dont les briques de bases dont des unitées CO_3 et CO_4 qui constituent des réseaux ou chaines plus ou moins complexes, plutôt que la molécule de CO_2 (figure 2).

En dessous de 10 GPa, la seule phase cristalline stable est la glace carbonique ("dry ice") CO_2-I [11, 16], qui cohabite avec une phase liquide à plus haute température [45]. Entre 10 et 13 GPa, comprimée dans une enclume a diamant, cette phase peut se transformer en trois phase en fonction de la température: en dessous de 450 K, elle se transforme en phase III [16], entre 450 et 600 K en phase IV [24] et entre 600 K et 800 K en phase VII [24] (au-dessus de 800 K elle fond [45]). L'ensemble de ces transitions sont soumises à des effets d'hysteresis [24], la transition I-III en étant l'exemple le plus important, la phase III pouvant n'apparaître qu'aux alentours de 18GPa sous compression de la phase I, et à l'inverse cette phase peut rester metastable jusqu'à 5 GPa sous décompression [16].

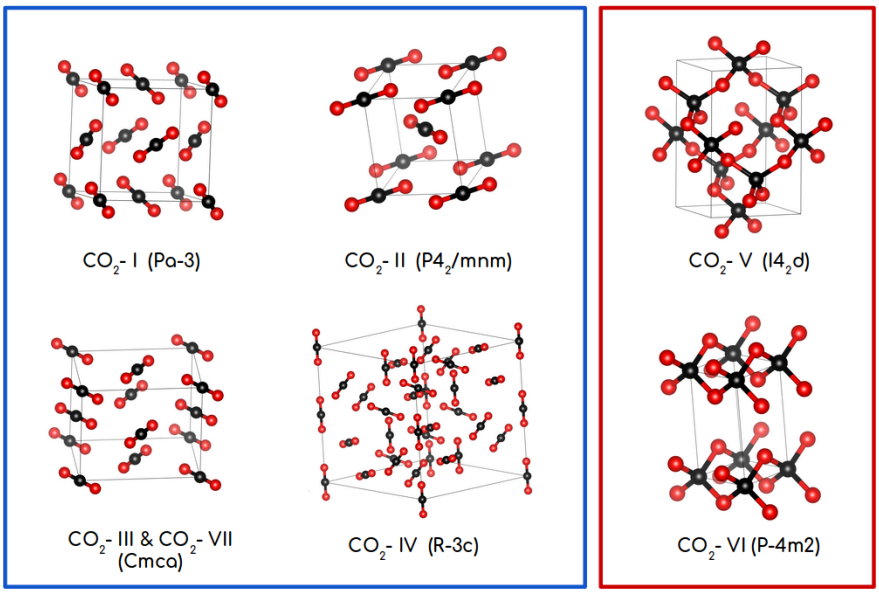


Figure 3: Phases cristallines du dioxyde de carbone sous haute pression.

La seconde zone moléculaire qui se trouve entre 10 et 40 GPa contient quatre phases cristallines moléculaires: II, III, IV et VII (figure 2). La phase II est une phase qui s'obtient à partir de la phase III [18] par recuit au dessus d'une température critique qui dépends de la pression. La phase IV [23] est une phase exisant au-dessus de 600K (figure 2), qui peut être obtenue en chauffant les phases II et III au dessus de 600 K. La phase VII quant à elle, ne peut s'obtenir que par compression de la phase I ou de la phase liquide [45]. En effet, même si sa zone de stabilité est bordée par la phase IV et que sous compression, on peut observer une transition VII-IV, la transformation IV-VII n'est pas observée experimentalement [24]. Il est d'ailleurs possible que les phases VII et III, qui partagent la même structure (à la différence d'un cellule plus allongée selon un axe pour la phase VII) soient en fait une seule et même structure, bien que leur région de stabilité soient disjointes [24, 46]. Ces deux structures seraient alors des intermediaires privilégiés

par des effets cinétiques entre la phase I et les phases II et IV (respectivement). Il est aussi notable que le comportement du liquide moléculaire évoluait sous compression dans la zone de pression correspondant à la transition I-III [6], ce qui indique également que la phase liquide moléculaire aussi peut aussi révéler des comportements variés et changeant avec les conditions expérimentales.

Au dessus de 40 GPa, les molécules de dioxyde de carbone réagissent pour former des phases polymérique, avec une pression de transition largement dépendante de la température. Les phases polymériques crystallines observées sont composées d'unités tetrahèdriques CO_4 [27, 47] reliées entre elles par des liaisons covalentes. À l'heure actuelle seules deux phases cristallines ont pu être observée experimentalement de façon reproductibles, les phases V [31, 32] et VI [34, 35, 36]. La phase V est une phase généralement obtenue par chauffe et compression des phases II, III et IV. Une fois formée, la phase est metastable jusqu'à très basse pression (au moins 5 GPa) [32, 48], et des expériences récentes ont également pu la retrouver à de très hautes pressions et températures (100 GPa et 25000 K) [49]. La phase VI elle se forme à partir des phases II et III à plus basse température que la phase V [33], et plusieurs études théoriques suggèrent une structure en couche[34, 36]. À plus basse température, la phase III comprimée au dessus de 60 GPa peut se transformer en phase amorphe [38, 39, 40, 50], aussi appelée a-carbonia, composée d'un mélange d'unités CO_3 et CO_4 formant des réseaux complexes. Enfin, au-dessus de 2000 K, trois scénarios sont envisagés: une dissociation du dioxyde de carbone [8, 9] (avec formation de diamant); une transformation du liquide moléculaire en liquide polymérique [7]; une soldification de la phase CO_2-V [37, 49]. Ces scénarios ne sont pas nécéssairement mutuellement exclusifs, l'importance des chemins thermodynamiques empruntés pouvant vraisemblablement avoir une influence sur le comportement obtenu.

Si l'ensemble des structures de ces différentes phases et leur zone de stabilités font l'objet d'un consensus, il reste que les transitions entre elles et les mécanismes associés sont toujours assez mal compris. De plus, certaines limites entre les différentes phases sont déterminées théoriquement en utilisant différences d'enthalpie ou d'energie libres entre les différentes phases [7, 37] et assez peu d'études théoriques se sont penchées sur les aspects cinétiques de ces transformations (à quelques exceptions près [35, 51]), qui peuvent pourtant être important [16]. Ainsi, si la transition entre la phase I et les phase III et VII sont martensitiques ne requierent qu'une rotation des molécules dans un plan, les transitions entre les phases II, III et IV ont été assez peu étudiées théoriquement et leur mécanismes et les aspects cinétiques sous-jacent sont encore mal connus. De même, la transition entre la phase moléculaire liquide et la potentielle phase liquide polymérique ou la phase cristalline polymérique V n'a été étudié que du point de vue energétique et reste très difficile d'accès pour les expériences.

De fait, dans cette thèse nous avons décidé de nous concentrer sur deux régions particulières du diagramme de phase du dioxyde de carbone:

- dans un premier temps, nous avons étudié la transition de phase entre les phases liquide moléculaire et liquide polymérique du dioxyde de carbone. Cette transition intervient dans des conditions de pressions et de température proches de celles du

géotherme du manteau inférieur et revêt donc un interêt certains pour une meilleure compréhension du manteau terrestre;

- dans un second temps, nous avons commencé une étude des transtions de phases entre les différentes phases moléculaires cristallines du dioxyde de carbone, afin de mieux en comprendre les mécanismes et les effets cinétiques associés.

Le plan de cette thèse est le suivant: dans la partie I, nous présentons l'ensemble des méthodes de simulations et d'analyse que nous avons utilisées dans l'ensemble de nos travaux; dans la partie II nous présentons des résultats liées aux deux objectifs de la thèse; la partie III, quant à elle, présente des résultats de travaux sur des sujets annexes, réalisés en parallèle. Enfin, nous présentons dans les appendices un ensemble de résultats d'analyses qui, bien qu'intéréssant, ne présentent pas assez de matière pour être présentés dans un chapitre à part entière.

Dans un premier temps, nous introduisons donc les méthodes de calculs qui ont permis la réalisation des simulations que nous avons utilisées pour ce projet. Nous présentons tout d'abord le principe des calculs *ab initio*, notamment via la théorie de la fonctionnelle de la densité électronique. Nous avons choisi ici de présenter ces méthodes de façon brève, ces éléments étant par ailleurs très bien décrits dans d'autres thèses, notamment celles de Félix Mouhat [52] et d'Adrien Mafety [53], pour ne citer qu'eux. L'intérêt de ces calculs est de permettre le calcul de l'énergie (et les propriétés associées) d'un système atomique en utilisant uniquement les positions des atomes dans une boîte de simulation au prix d'un coût numérique relativement modéré. Nous introduisons aussi la dynamique moléculaire, qui permet d'intégrer les mouvements des atomes afin de calculer des moyennes thermodynamiques et d'observer les mécanismes des réactions chimiques et/ou mécanismes de transition de phase.

Nous détaillons, dans le chapitre suivant, différentes méthodes permettant de décrire numériquement un système à l'aide de variables numériques, que nous appellerons de façon interchangeable descripteurs ou variables collectives. Ces variables sont utilisées soit afin de permetre l'exploration des paysages d'énergie libre, soit dans l'analyse des données de simulation. Nous présentons ces variables en deux groupes selon qu'elle décrivent un système au complet (variables globales) ou un environnement atomique spéficique (variables locales). Parmi ces variables nous insistons particulièrement sur le vecteur invariant par permutation (ou PIV [54]) et sur le variables sociales invariantes par permutation (ou SPRINT [55]) qui seront toutes deux particulièrement utilisées dans cette thèse. Tant que possible, nous décrivons les aspects computationnels liés à ces variables ainsi que leur potentielles limitations.

Le troisième chapitre de la partie méthode est dédié aux méthodes d'apprentissage statistiques (aussi connues sous le nom de *machine learning* (ML)) que nous avons pu utilisés ou conçues lors de cette thèse. Nous décrivons notamment un ensemble de méthodes de classification non-supervisées qui seront utilisées par la suite afin de regrouper des structures similaires en utilisant les variables collectives mentionnées précédemment. Nous introduisons ensuite les méthodes des états de Markov, qui permettent de modéliser

l'évolution temporelle de systèmes evoluant entre différents états discrets de façon aléatoire et sans mémoire et qui sera mise en application pour l'étude de la dynamique chimique des atomes dans le liquide polymérique haute température par la suite.

Enfin, le dernier chapitre de la partie méthode décrit des méthodes d'exploration de paysages d'energie libre, que nous divisons en fonction de leur applications: l'analyse des transformations de la matière avec les méthodes de métadynamique et d'échantillonage par ombrelles ou bien la recherche de structures stables à partir d'une configuration chimique donnée.

Nous présentons dans le premier chapitre de la partie consacrée aux résultats, ceux que nous avons obtenues sur le principal sujet de cette thèse: la transition entre une phase moléculaire liquide et une phase polymérique liquide dans les conditions géothermiques.

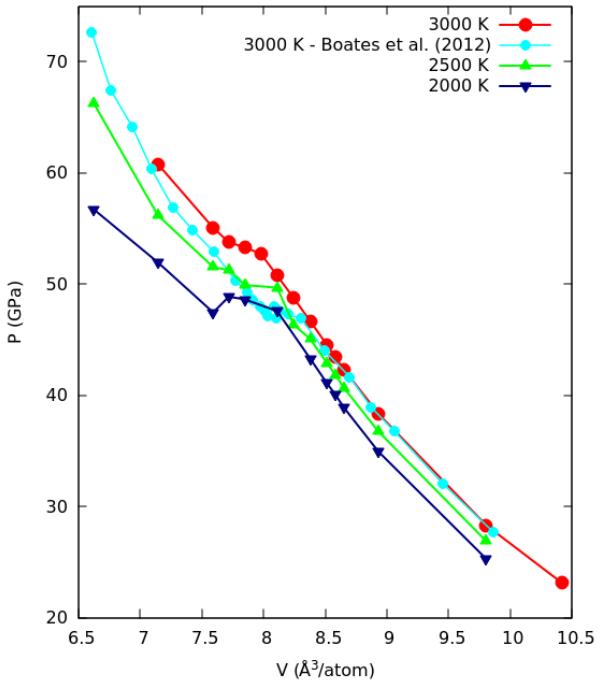


Figure 4: Pression en fonction du volume de la boîte de simulation à 2000, 2500 et 3000 K, avec comparaison avec les résultats obtenus par Boates et al. (2012) [7] à 3000 K.

Afin d'étudier le comportement du liquide dans ces conditions, nous avons utilisé des simulation de dynamique moléculaire *ab initio* avec des temps de productions longs (100ps) afin d'avoir suffisament de données pour avoir un bon echantillonage statistique. Ce faisant, nous avons étudié le système à des pressions entre 30 et 70 GPa et trois températures différentes: 2000, 2500 et 3000K.

Nous étudions dans un premier temps l'évolution de la pression exercée sur le système en fonction du volume de la cellule de simulation (figure 4). Cette relation permet de

mettre en évidence (figure 4), comme chez Boates et al. [7], un région plateau, qui indique une transition du premier ordre entre les deux phases liquides.

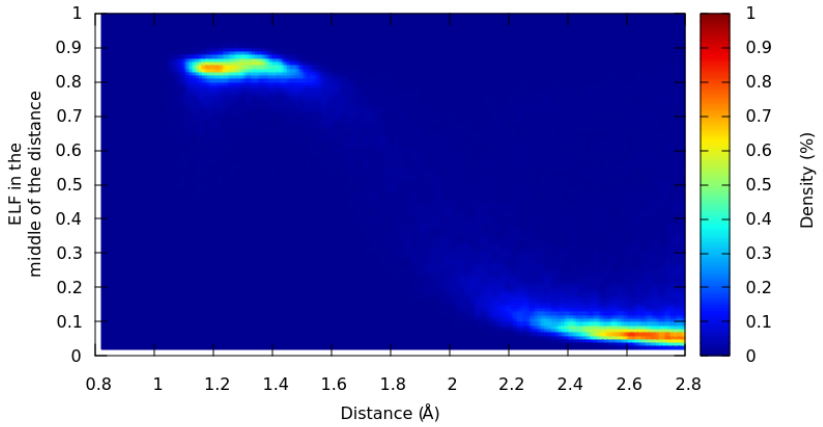


Figure 5: Distribution des valeurs de la distance interatomique et de fonction de localisation électronique (ELF) au milieu de cette distance. Une liaison atomique est caractérisée par des valeurs de la ELF supérieures à 0.7 et semble corrélée fortement à des faibles distances interatomiques.

De façon à analyser en profondeur les méchanismes chimiques se produisant dans le liquide polymérique, nous avons tout d'abord réalisé un ensemble d'analyses permettant d'établir la validité de l'utilisation de valeurs seuils sur les distance pour determiner les liaisons entre atomes de carbone et d'oxygène. Pour ce faire nous utilisons les fonctions de localisation électroniques (ELF) [56] qui permettent d'obtenir un critère de validation basée sur la densité électronique (figure 5), ainsi qu'une approche basées sur une méthode de classification [57] qui apporte un critère statistique prenant en compte l'environnement local des atomes.

Une fois les valeurs seuils validées, nous pouvons les utiliser pour mettre en évidence la transition liquide-liquide en calculant la fraction des carbones avec coordinance 2, 3 et 4 voisins en fonction de la température et de la pression. Nous observons que les carbones avec deux voisins deviennent soudainement moins nombreux autour de la pression de transition (située entre 48 et 55GPa en fonction de la température). Cette diminution correspond à un remplacement des unités CO_2 en unités CO_3 et CO_4, les premiers étant plus fréquents à 3000K et les second à 2000K, ce qui donne une première indication d'une différence de comportement entre le fluide polymérique à 2000 et 3000 K.

La différence de coordinance entre les liquides à haute et basse température se repercute aussi sur leurs propriétés dynamiques et structurales. Nous observons, en effet, une réduction très forte du coefficient de diffusion à basse température, suggérant une possible amorphisation. De plus, la fonction de corrélation de paire (figure 7) entre les atomes de carbone et d'oxygène renforce cette interprétation, avec des pics secondaires bien mieux

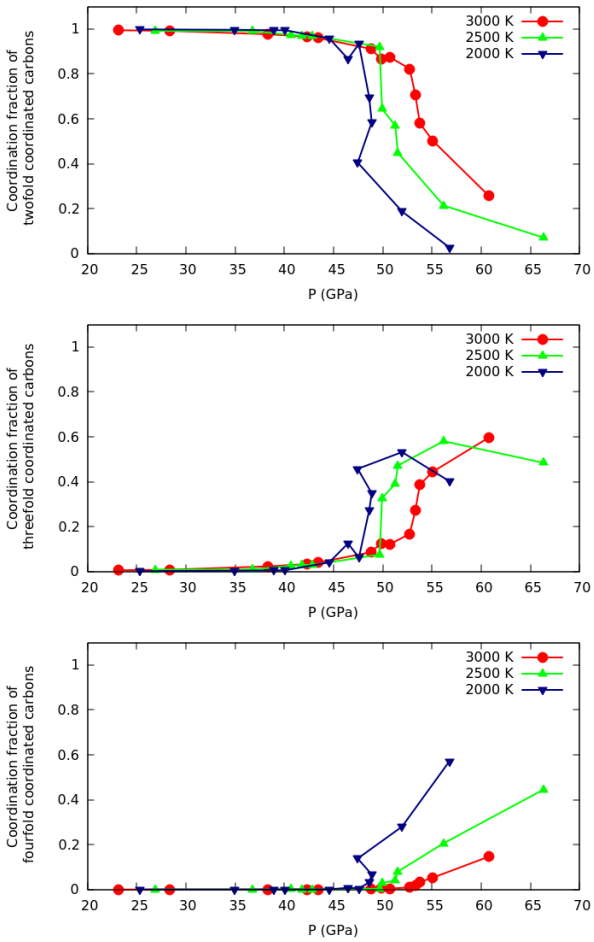


Figure 6: Fraction des coordinances des atomes de carbone en fonction de la température et de la pression. Les carbones avec deux voisins sont en haut, ceux avec trois voisins sont au milieu et les carbones avec quatre voisins en bas.

définis à 2000 K qu'à 3000 K.

En utilisant la distribution des tailles des molécules, nous observons que le liquide polymerise également à des rythmes différents à ces deux températures: la polymérisation est brusque à 2000K où elle devient totale dès 55 GPa alors qu'elle est très progressive à 3000 K, avec une stabilisation progressive de chaînes de tailles intermédiaires et où la polymérisation ne commence à être totale que vers 65 GPa.

Nous constatons donc deux comportements polymériques très différent en fonction de la température: un liquide polymérique très réactif à haute température (3000 K) avec une formation majoritaire d'unités CO_3, et un liquide polymérique presque amorphe à 2000 K, dominé par des unités CO_4 formant un réseau tridimensionnel complexe.

En analysant le liquide moléculaire, nous avons également mis en évidence un autre type de comportement fluide: en effet, dans les régions à haute température ($T > 2500K$) et/ou pression ($P > 40$ GPa) le liquide moléculaire devient réactif, et on observe la for-

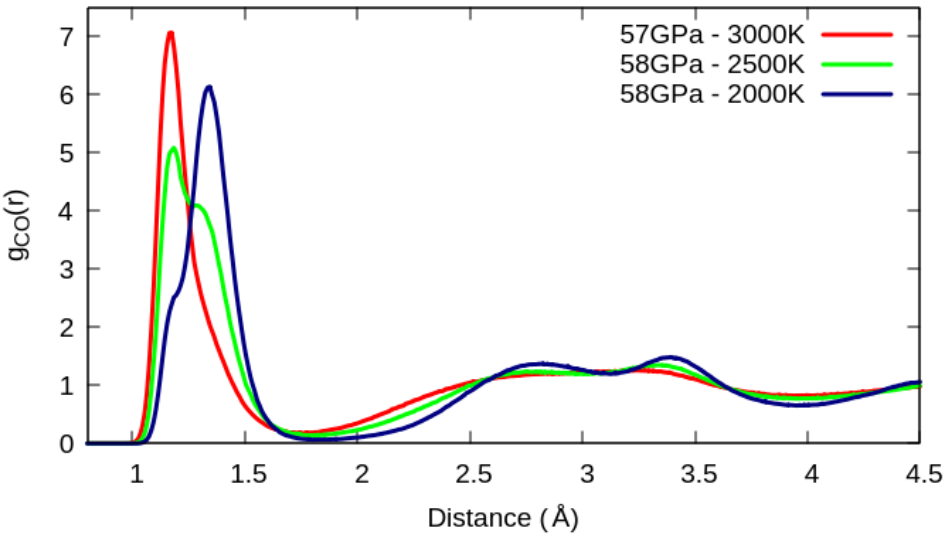


Figure 7: Fonction de corrélation de paires entre les atomes de carbone et d'oxygène autour de 55 GPa à 2000, 2500 et 3000 K.

mation régulière de courtes chaines C_2O_4, formant occasionnellement des dimères qui permettent des échanges de deux des oxygènes des molécules en intéraction (figure 8). Ces molécules avaient déjà été observées à 4000K [58], et dans des phases amorphes autour de 150 K[59], mais elles apparaîssent ici fréquemment et semblent caractéristique d'un comportement liquide particulièrement réactif, tranchant avec la description usuelle du liquide moléculaire où les dioxyde de carbone n'interargissent entre elles que par des interactions faibles. Ce comportement est d'autant plus intéréssant qu'il survient dans la zone de formation potentielle du CO_2 en profondeur [4, 5], suggérant que le dioxyde de carbone pourrait ainsi prendre une part très active à la chimie du manteau inférieur de la Terre.

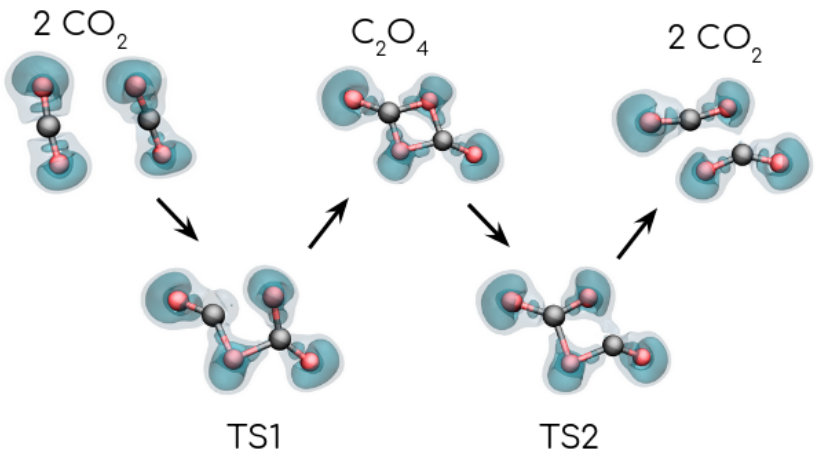


Figure 8: Mécanismes de la dimérisation avec échange d'oxygène entre deux molécules de CO_2 avec les isovaleurs de ELF (bleu: 0.8, noir 0.6).

Nous terminons cette étude sur les liquides polymériques en utilisant des états de markov [60] afin de modéliser l'évolution des atomes de carbone au cours d'une simulation. En étandant la description du voisinage atomique jusqu'au second voisins, nous arrivons à établir des états atomiques (la molécule de CO_2, unités CO_3, CO_3-, etc...), entre lesquels la dynamique des atomes de carbone markovienne. Cette analyse nous permet d'avoir accès à la dynamique individuelle des atomes, et donc potentiellement à la réactivité des atomes sur l'ensemble du champ de pression et de température où cette analyse est appliquable.

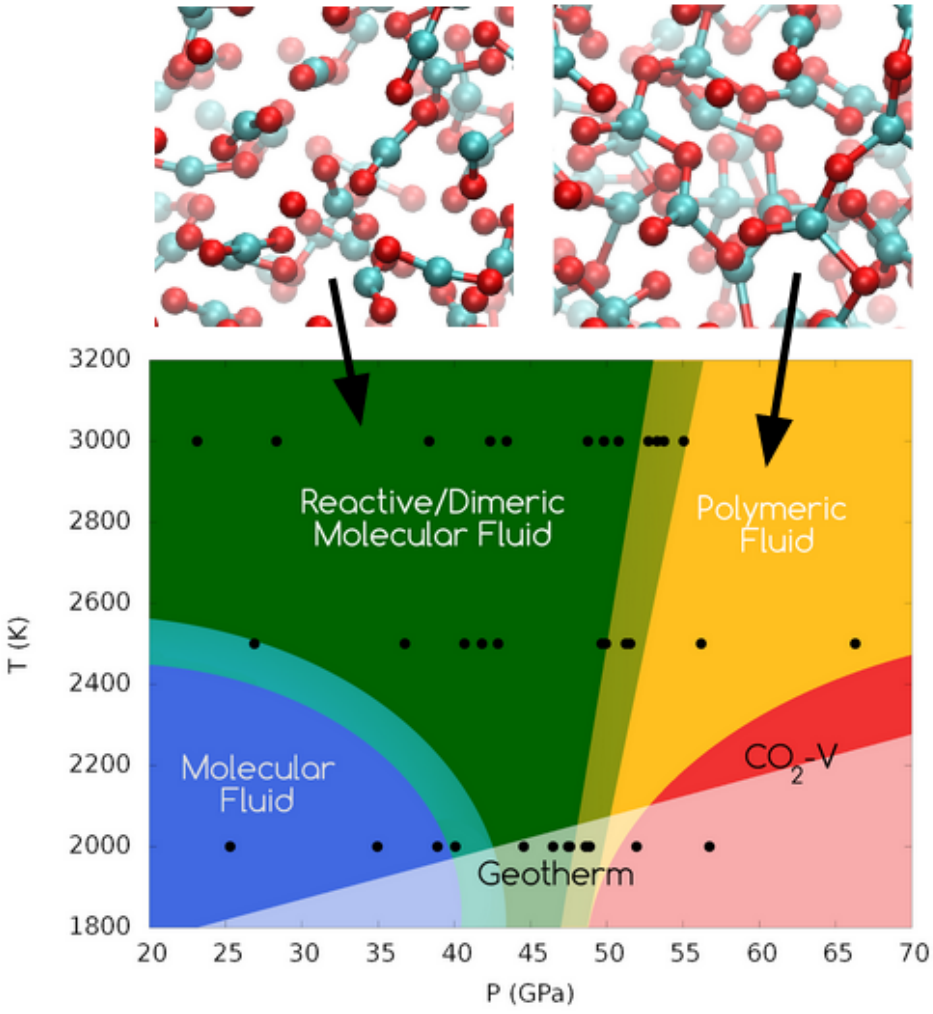


Figure 9: Diagramme de phase tel que proposé suite à nos analyses dans la région de transition entre fluide moléculaire (en haut à gauche) et fluide polymérique (en haut à droite).

Au final, cette étude a mobilisé des outils d'analyse structurale et dynamique poussés et a permis la mise en évidence de quatre comportements fluides distincts (figure 9): un liquide moléculaire simple où les molécules de dioxyde de carbone n'interargissent que par interactions faibles; un liquide moléculaire réactif avec formation de dimères; un liquide polymérique dynamique à forte réactivité; un liquide polymérique potentiellement amor-

phe.

Dans un deuxième chapitre, nous présentons deux études sur les phases cristallines du dioxyde de carbone: la première consistant en l'utilisation d'une recherche des structures stables du dioxyde de carbone à haute pression avec la méthode AIRSS [61]; la seconde en analyse des transitions entre les diffèrentes phases cristallines moléculaires du dioxyde de carbone et notamment entre la phase I et et la phase III.

L'utilisation méthode de recherche de structure AIRSS au dioxyde de carbon nous a permis de tester la méthode sur un système présentant un changement important de type de structure avec la pression, et s'inscrivait dans la logique d'une transmission de connaissance sur cette méthode à l'intérieur de l'équipe. À l'exception de la phase CO_2-IV (dont le nombre d'unités CO_2 par maille unitaire est trop grand pour que la méthode de recherche puisse la retrouver à un coût de calcul modéré), nous retrouvons l'ensemble des structures mentionnées précédement et nous présentons un diagramme de phase à 0 K de l'ensemble de ces phases basés sur des calculs *ab initio*.

Nous introduisons une analyse de la transition entre les phases I et III au moyen de la métadynamique et du vecteur invariant par permutation (PIV) [54]. Nous arrivons à reproduire les résultats obtenus précedemment par Gimondi et al. [51] en utilisant le même champ de force.

Nous présentons ensuite, dans une partie liée à des projets satellites, une méthode pour explorer l'espace des configurations de nano-clusters et de petites molécules de façon à trouver l'ensemble des structures stables associées à une configuration chimique de manière non supervisée. Pour ce faire cette méthode se base sur l'utilisation de dynamique moléculaire *ab initio* accélérée par métadynamique afin de pousser le système à pousser le système à évoluer en dehors des configurations déjà explorées. Cette méthode permet ainsi de visiter progressivement l'ensemble des structures accessibles par le système. Nous utilisons ensuite un algorithme de classification non-supervisée pour faire une partition de l'espace des configurations. Les centres de cette partition sont optimisés géométriquement avec des calculs *ab initio* poussés et les structures stabilisées constituent les structures stables détectées.

Nous appliquons cette méthode à des clusters de MoS_2 de trois tailles différentes (Mo_2S_4, Mo_3S_6 et Mo_4S_6) et trouvons un ensemble de plus de cent structures différentes, dont un grand nombre n'avait pas été reporté précédemment. Pour chacune de ces structures, nous calculons l'énergie de liaison, la différence d'énergie entre la plus haute orbitale non-occupée et la plus basse orbitale non-occupée et la magnétisation.

La méthode que nous présentons remplit dès lors ses objectifs, et peut être appliquée à des systèmes similaires, comme des petits clusters ou des molécules. Elle est cependant limitée par le coût potentiellement fort des longues trajectoires métadynamiques *ab initio*, ce qui la rends difficilement appliquable pour de plus grandes structures. Ce projet a été réalisé avec la collaboration de Sofiane Schaack, doctorant à l'INSP, et a été soumis dans

la revue *The Journal of Chemical Physics C* sous le titre **"Unsupervised computer exploration of MoS_2 nanoclusters: structures, energetics, and electronic properties"**.

Nous présentons ensuite les résultats d'un travail en coopération avec un post-doctorant de l'équipe, Gabriele Mogni, sur l'accelération de la méthode de recherche AIRSS [61]. Dans ce travail nous montrons qu'il est potentiellement plus rapide de faire un scan préalable de l'ensemble de l'espace des configurations plutôt que de faire des optimisations successives à des points aléatoires de l'espace pour récupérer les minimum locaux de l'espace des configurations. Nous proposons également une méthode permettant d'évaluer la convergence de la recherche de structure en nous basant sur une méthode de classification non-supervisée [62].

Nous présentons en appendice quatre études courtes sur des sujets divers: nous présentons d'abord des résultats d'une analyse spectroscopique de la phase V et d'une potentielle nouvelle phase non-moléculaire proposée par Yong et al [48]. Nous présentons ensuite une continuation de l'étude sur l'utilisation de la valeur de ELF au milieu des distances interatomiques pour déterminer les liaisons marche sur d'autres systèmes que le CO_2 avec une étude sur un système contenant de la glycine solvatée dans de l'eau, en montrant notamment que cette méthode semble être également efficace sur ce système pourtant très diffèrent. L'appendice suivant présente des tests du vecteur invariant par permutation qui permettent de montrer qu'il n'existe en pratique pas de grande différence d'énergie entre deux structures proches dans l'espace du PIV, ce qui valide au moins partiellement ce descripteur; enfin, nous présentons un courte étude spectroscopique de la phase Ima2 de NH_3-H_2O, en continuité d'une étude de Adrien Mafety [53].

Dans cette thèse, nous établissons donc les grands traits d'un diagramme de phase complexe et riche du dioxyde de carbone en conditions extrême. Nous montrons notamment des résultats indiquant l'existence d'un liquide moléculaire réactif dans des régions d'interêt pour la compréhension du manteau profond. Cette thèse a aussi permis l'établissement de plusieurs méthodes d'analysis basées sur les descriptions des systèmes, l'utilisation de critères électroniques mais aussi de méthodes d'apprentissage statistique.

Contents

List of Figures

List of Abbreviations

CM	Coulomb Matri
CV	Collective Variable
DFT	Density Functional Theory
DPC	Density Peak Clustering
ELF	Electron Localization Function
FES	Free Energy Surface
fu	formula unit
fs	femtosecond(s)
GGA	General Gradient Approximation
GPA	GigaPascal
LDA	Local Density Approximation
MSD	Mean Square Distance
PP	PseudoPotential(s)
ps	picosecond(s)
SPRINT	Social PeRmutation InvariaNT
VdW	Van der Waals

Introduction

Carbon dioxide is one of the most important chemical species in nowadays life, at the heart of a massive research effort, particularly in geochemistry[1, 2], atmospheric chemistry and climate science, due to its importance in global warming. It is indeed a simple and abundant molecular system, that can relatively easily crystallize into dry ice CO_2-I. It has also the particular property of interacting mainly through quadrupolar interactions, which made it a unique toy model to understand the behavior of molecular crystals made of weakly interacting chemical units.

Of course, even before its impact on climate was widely accepted by the scientific community[63, 64], CO_2 was already the object of extensive research, particularly at high pressure conditions [10, 11, 12], both, as mentioned above, as an emblematic fundamental molecular system, and as an important constituent of the Earth interiors, in the fluid phase and within carbonate minerals.

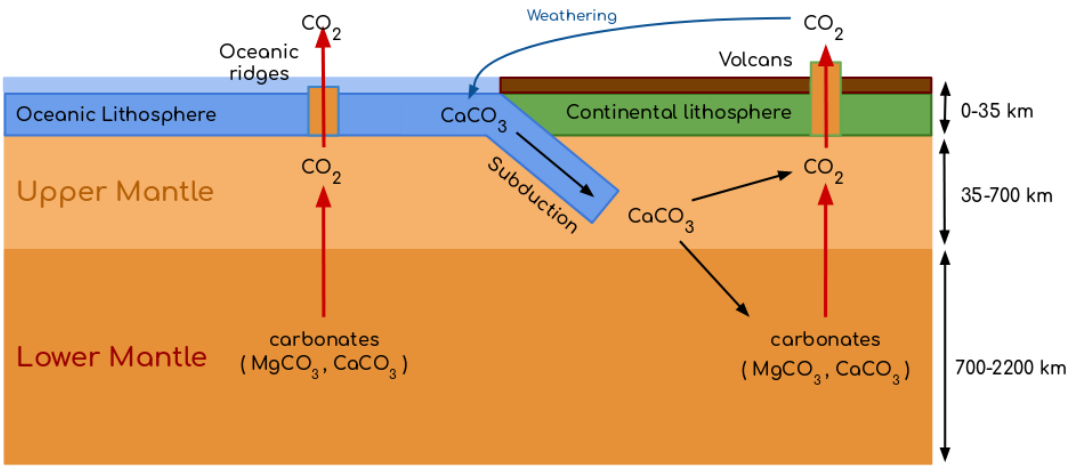


Figure 10: The deep carbon cycle

Carbon dioxide is in fact present in the Earth mantle [1, 2, 3] and its contribution to geophysical phenomena such as volcanism and earthquakes has been extensively investigated [1, 2]. Its presence in the depth of the Earth makes it an important part of the carbon cycle [3] (figure 6.3). It is formed as the result of various chemical reactions between the carbonates of the mantle and its constituents; these carbonates themselves are buried into the mantle through the subduction of crusts containing organic sediments.

Finally, the emission of carbon dioxide through volcanism completes the deep carbon cycle by releasing CO_2 into the atmosphere [3]. Although the presence of carbon dioxide in the mantle is more generally associated with the upper mantle, it has been recently suggested that carbon dioxide could form in the depths of the lower mantle through reactions of the carbonates (such as $MgCO_3$ and $CaCO_3$) and SiO_2 [4, 5].

In this context, it is not surprising that carbon dioxide has been frequently the object of intensive studies in the high pressure field [13, 17, 18, 20, 21, 22, 23, 65]. A rich panel of crystalline phases have been found experimentally, whose structure have remained quite controversial for at least a decade. Over the years, the combination of experimental and theoretical work proved complementary [6, 32, 38, 39, 40, 31] and made it possible to settle most of the debates about crystalline high-pressure carbon dioxide.

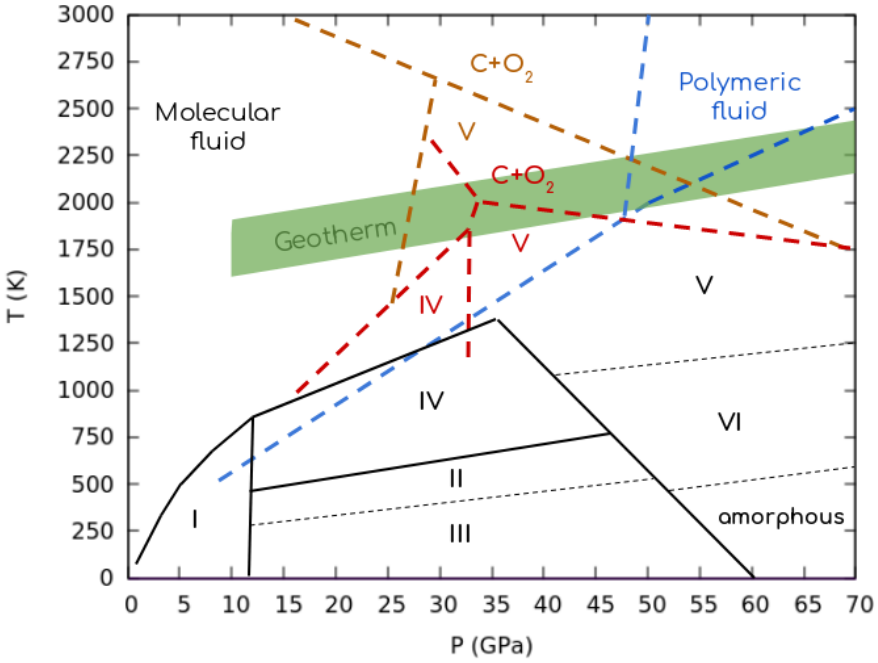


Figure 11: High pressure phase diagram of carbon dioxide, using results from [6]. **Blue:** melting line of carbon dioxide and transition limit between molecular and polymeric liquids according to [7]. **Red:** limits between phase IV, V, molecular fluid and dissociation of carbon dioxide according to [8]. **Brown:** limits between phase V, molecular fuid and carbo dioxide dissociation as reported in [9]. Dotted lines indicate kinetic lines.

To give a broad, general picture of the high pressure phase diagram of CO_2, one can roughly be separate it into two region: the first between 1 and 40 GPa and the other above 40 GPa (figure 11). In the 1-40 GPa range, carbon dioxide phases are all molecular, either crystalline or fluid (figure 11). It exhibits up to 5 different molecular crystal phases (labelled CO_2-I, II, III, IV, and VII, figure 12), while the carbon dioxide fluid was shown to exhibit two different structural behavior [6]. In particular, in this low-to-moderate part of the phase diagram, one observes CO_2-I up to roughly 12 GPa [11, 16],

which then transforms upon compression into either CO_2-III[16], CO_2-VII [24] or CO_2-IV [20, 21, 22, 23] depending on the temperature. Structural similarities between CO_2-III and CO_2-VII, despite their relative distance in the phase diagram, have long hinted that those phases are actually the same, although their respective stability zone are disjoint [24, 46].

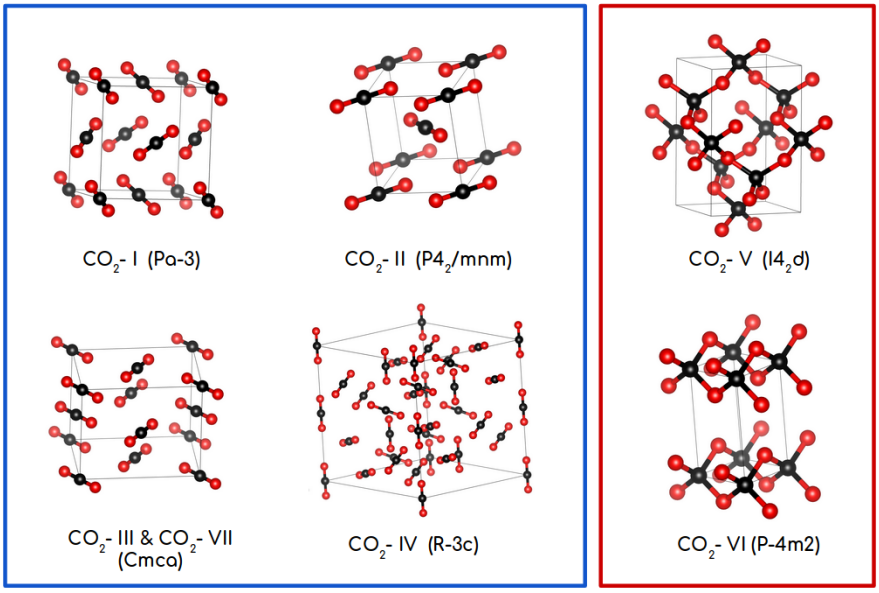


Figure 12: Crystalline phases of high pressure carbon dioxide.

CO_2-III and CO_2-VII transform into CO_2-IV [24] if heated. CO_2-II can be obtained through annealing of CO_2-III [17, 18], and it may in turn transform into CO_2-IV if heated [22, 23]. Phases I, VII and IV can be melted into a molecular liquid whose properties continuously change upon compression, somehow mirroring the transformation of CO_2-I into CO_2-III [6]. Although there is a consensus on the structures of all those phases, the transitions between them are by no means simple, their mechanisms remain elusive, and frequently affected by important metastability and hysteresis effects.

Above 40 GPa, phases II, III and IV transform into an array of polymeric phases whose nature depends heavily on the thermodynamic path (figure 12). CO_2-III can transform either into an amorphous phase at low temperature [38, 39, 40], but also upon eating into either of two polymeric crystal phases, CO_2-V [66, 31, 32, 67] and CO_2-VI. CO_2-II and -IV have been observed to also transform into CO_2-V [8].

The high temperature behavior of carbon dioxide of this part of the phase diagram has proven more complex to explore for experimentalist and the contributions of theoretical studies are therefore all the more valuable. At temperatures above 2000 K, experiments are difficult and their results contradictory: although in two separate instances dissociation of carbon dioxide was observed [8, 9] at 2000K around 40 GPa, CO_2-V was recently recovered at 100 GPa and 2500 K [49]. On the other hand, theoretical studies focusing on the compression of the molecular fluid have shown evidences pointing to a first order

liquid-liquid phase transition into a polymeric fluid with large carbonate chains, and/or into polymeric crystal phase V [7], while subsequent analysis based on energetics considerations hinted toward solidification into phase V [37]. Interestingly, theoretical calculations on the high temperature range of the molecular liquid showed evidence of dimerization [58], indication of an unsuspected reactive behavior of carbon dioxide.

In this work we first investigate carbon dioxide fluids behavior under geological conditions in order to study in details the transformation of the molecular liquid into a polymeric fluid. In so doing, we use a variety of advanced methods to analyze the system, including methods based on the Electron Localization Function [56] and Density Peak Clustering [62] to analyze bonds between atoms, but also Markov State Models[60] to analyze the chemical dynamics of the atoms. The second objective of this work is the analysis of the phase transitions between molecular phases of carbon dioxide using enhanced sampling methods and advanced descritptors such as the Permutation Invariant Vector (PIV).

Although this thesis is mainly – but not exclusively – devoted to the study of carbon dioxide at extreme conditions of pressure and temperature, it must be underlined that our research group has undergone, precisely in the last three years, a significant change of philosophy and approach in the general study of transformations in condensed matter and chemical systems. In particular, strong efforts have been devoted to the development of novel analytical tools, capable to efficiently infer the deep structural and dynamical properties of a given system from the topological/statistical features of their atomic configurations in space and time. As detailed in chapter 3, those efforts were focused on clustering methods, data analysis, and machine learning approaches. Although those tools were decisively useful only in a fraction of the cases, on the understanding of the behavior of high-pressure CO_2, the latter has regularly been for us an ideal "test-bed", because of the large accumulated trajectories and data, and of its nature of a homogeneous system, but undergoing several modifications.

The plan of this thesis is as follow:

- In the first part of this work, we introduce all the methods that we have used during the course of this thesis both for calculations and analysis purposes.
- In part II we focus on our results on carbon dioxide, first focusing on the results obtained on the liquid-liquid phase transition at geological pressure. We provide evidence for the existence of four distinctive fluid behaviors, all withing the bounds of the geothermal conditions corresponding to depth where carbon dioxide may form [4, 5]. We then present a series of short work on crystalline molecular phases of carbon dioxide: we show that the AIRSS methods is able to recover most of the stable crystal phases of carbon dioxide under high pressure; finally using metadynamics and the permutation invariant vector, we show encouraging preliminary results for the study of transitions between molecular phases of CO_2 obtaining the transition between phase I and III at 5GPa and 3000K.
- In part III we describe two satellite projects which we carried out in parallel with this thesis, using similar tool as in our main work. We first showcase a method to

explore the configuration space of small molecules and/or clusters using a mix of metadynamics, unsupervised learning and social collective variables. We show the effectiveness of the methods on small MoS_2 nanoclusters. We then report on participation to a joint projet with a former post-doctoral researcher, Gabriele Mogni, that focused on accelerating a search of structure algorithm, and propose a new criterion method to test the progress of a structure search method.

- In the appendices we showcase a few interesting small analysis and side projects that did not show enough results to deserve a separate chapter.

Part I

Methods

Chapter 1

Simulations in condensed matter

Introduction

At the begining of the 20^{th} century, with the emergence of quantum mechanics, one could hope that the understanding of condensed matter was essentially within reach, as the fundamental equations that described its behavior were completely known. However, it turned out that the Schrödinger equation proved to be impossible to solve analytically for systems more complicated than the hydrogen atom, in a twist reminiscent of the N-body problem for massive bodies for astrophysics. However, while it proved feasible to solve numerically the N-body problem without making more approximation to the physical model, even the numerical resolution of the complete Schrödinger equation is beyond reach except for some very simple cases.

In this section we introduce a powerful solution to that limitation: *ab initio* calculations, which is a large set of frameworks making careful approximations on the original Schrödinger equation so that its numerical resolution is accessible on modern computers.

1.1 *Ab Initio* calculations

1.1.1 The Born-Oppenheimer approximation

We start with the original hamiltonian for a system of n_N nuclei and n_e electrons in interactions:

$$H = T_N + V_{N,N} + T_e + V_{e,e} + V_{e,N} \tag{1.1}$$

Where T_N is the nuclear kinetic energy operator:

$$T_N = -\frac{\hbar^2}{2}\sum_{i=1}^{n_N}\frac{\nabla_i^2}{M_i} \tag{1.2}$$

T_e is the corresponding electronic kinetic operator:

$$T_e = -\frac{\hbar^2}{2m_e}\sum_{j=1}^{n_e}\nabla_j^2 \tag{1.3}$$

$V_{e,N}$ is the coulomb interaction between nuclei and electrons:

$$V_{e,N} = -\sum_{i=1}^{n_N}\sum_{j=1}^{n_e}\frac{Z_i\ e^2}{|r_j - R_i|} \tag{1.4}$$

the coulomb interaction between nuclei are represented by $V_{N,N}$:

$$V_{N,N} = \sum_{i=1}^{n_N-1}\sum_{i'=i+1}^{n_e}\frac{Z_i Z_{i'}\ e^2}{|R_i - R_i'|} \tag{1.5}$$

and the electron-electron coulomb interaction is capture in $V_{e,e}$:

$$V_{e,e} = \sum_{j=1}^{n_e-1}\sum_{j'=j+1}^{n_e}\frac{e^2}{|r_j - r_j'|} \tag{1.6}$$

In this equations, i and i' sum over the nuclei, j and j' sum over the electrons, M_i is the mass of the ith nuclei and m_e is the electronic mass.

As we mentioned, although the system's behavior is in principle fully determined by this equation, it is in practice not solvable. In order to make predictions for more general systems, we use the Born-Oppenheimer approximation [68] in which we consider that the nuclei are much heavier than the electrons and therefore evolve over much longer timescales. In other words, there is in general little coupling between the behavior of the electron cloud and the nuclei, and it is reasonable to first solve the behavior of the electron cloud behavior,assuming the nuclei to be static, and then to compute the energy of the system by calculating the interactions between the electrons with the nuclei, in order to finally get the properties of the system.

We can therefore write the total ket of the system $|\phi(r,R)\rangle$ as a formal product of the electronic $|\psi(r,R)\rangle$ and nuclear $|\chi(R)\rangle$ ones:

$$|\phi(r,R)\rangle = |\psi(r,R)\rangle \otimes |\chi(R)\rangle \tag{1.7}$$

with r being the cartesian coordinates of the electrons while R are those of the nuclei.

Following up on this approximation, we assume that the electron cloud follows the nuclei adiabatically and we can write an electronic hamiltonian $H_e(R)$ that depends only on the positions of the nuclei:

$$H_e = V_{N,N} + T_e + V_{e,e} + V_{e,N} \tag{1.8}$$

Which we can solve in order to obtain the electronic wavefunction $H_e(R)|\psi\rangle = E_e(R)|\psi\rangle$ and energy $E_e(r)$ which can then be used in order to solve the nuclear hamiltonian H_N:

$$H_N = T_N + E_e(R) \tag{1.9}$$

We will call *ab initio* or "first principle" calculations all such scheme that solve numerically the electronic part of the Schrödinger equation, as they only require the position of the nuclei in a system, in order to be able to solve for its electronic density and related properties. The following section will cover a framework known as density functional theory that allows this to be done in practice at relatively moderate computational cost.

1.1.2 Density Functional Theory

We will only make a short description the density functional theory in this section, as the method itself was not at the core of the thesis. The interested reader may find more information in the theses of Adrien Mafety [53], Félix Mouhat [52] or in the reference document by Fabio Finocchi [69] which all have been sources of inspiration to write this section.

1.1.2.1 The theory

In order to solve the electronic hamiltonian (1.8), Density Functional Theory (DFT) takes a unique approach: instead of the wavefunction, it uses the electronic density as the central quantity. The reason behind this is that most of the physical properties that one can get from the electronic wavefunction can in principle be computed using the electronic density. Moreover the density is a real scalar field in 3 dimensions whereas the electronic wavefunction is a $3\,n_e$ complex field, which means its dimension will increase with the number of electrons in the system, making it ever more complex to solve. This elegant approach was proposed in the early 60's by P. Hohenberg, L.J. Sham and W. Kohn [70, 71], the latter becoming a Nobel Prize in Chemistry for this scientific contribution.

In the following we will note the electron density as ρ and define it as:

$$\rho(r) = n_e \int dr_2 ... dr_{n_e} |\psi(r_2, \, ... \, , r_{n_e})|^2 \tag{1.10}$$

With n_e the number of electrons in the system, and r_i denoting the cartesian coordinates of the i^{th} electron.

In order to solve for the ground state electronic density, density functional theory relies on two specific theoretical pieces: the Hohemberg and Kohn Theorem and the Kohn Sham equation.

The Hohemberg-Kohn Theorem [71] states that there is a one-to-one correspondence between the ground state hamiltonian and the ground state electronic density[1]. We can therefore write the energy as a functional of the electronic density (hence the name of the method), by taking a clue from (1.8):

$$E[\rho] = T[\rho] + E_{e,e}[\rho] + E_{ext} \tag{1.11}$$

where:

- $T[\rho]$ is the kinetic energy functional;
- $E_{e,e}[\rho]$ is electron-electron interaction functional;
- E_{ext} is "external" potential energy functional that depends only on the position of the nuclei;

The main issue with this expression is that there is no known analytical form both for $T[\rho]$ and $E_{e,e}[\rho]$. Regardless, from this expression of $E[\rho]$, the Hohenberg-Kohn theorem states that $E[\rho]$ is minimal when ρ is the actual ground state energy ρ_0, which means that we can use a variational principle to find the ground state energy through the electronic density:

$$\frac{\delta}{\delta n}\left[E[\rho(r)] - \mu \int d^3r\ \rho(r)\right] = 0 \tag{1.12}$$

where the second term has been introduced to maintain constant the number of electrons in the system, μ being a lagrangian multiplier.

In order to go further, one requires analytical expressions for both $T[\rho]$ and $E_{e,e}$, which is the central point of the Kohn-Sham equation. This method relies on the assumption that it is always possible to find, for any system composed of many interacting electrons, a corresponding virtual system of non interacting electrons (that is, where the electron-electron interaction is mediated by an effective "external" potential V_{eff}) with the same electronic density. That is:

$$\rho = n_e \sum_{i=1} |\phi_i(r)|^2 \tag{1.13}$$

[1] The proof of this theorem may be found either in the original article [71] or in a course notes of F. Finocchi [69]

where $\phi_i(r)$ is the wavefunction of the i^{th} virtual electron in the non-interacting system.

Building on this, the kinetic energy functional $T[n]$ may be expressed as that of the non-interacting system $T_s[\rho]$ whose analytical expression is known. The corresponding electron-electron density functional may be expressed as a simple coulomb interaction functional called Hartree term ($E_H[\rho]$) which can also be expressed analytically. Regrouping the interaction terms that are not taken into account by those term into a single term called the exchange and correlation functional $E_{xc}[\rho]$, we finally obtain the first part of the Kohn-Sham equations:

$$E[\rho] = T_s[\rho] + E_H[\rho] + E_{xc}[\rho] + E_{ext} \tag{1.14}$$

The $E_{xc}[\rho]$ term therefore regroups all the remaining unknown terms. It is composed of two kind of contributions: the effects from the quantum particles indistinguishability and corrections on the kinetic energy due to the interactions between electrons.

For the second part we simply need to write the Schrödinger equations for all the system of non interacting electrons:

$$\left[-\frac{\hbar^2}{2m_e} + V_{eff}(r) \right] \phi_i(r) \;=\; \epsilon_i \, \phi_i(r) \tag{1.15}$$

Where i refers to the i^{th} virtual non-interacting electron. In order to determine the expression of $V_{eff}(r)$ can be derived from $E[\rho]$ as:

$$V_{eff}(r) = V_{ext}(r) + e^2 \int dr' \frac{\rho(r')}{|r - r'|} + V_{xc}(r) \tag{1.16}$$

where $V_{xc}(r) = \frac{\delta E_{xc}}{\delta n}$ and the second term accounts for electronic repulsion between the electron and the electron cloud.

All the necessary ingredients are now present, and to solve one just needs to solve (1.14), (1.16) and (1.13) in a self consistent manner (figure 1.1): starting with a trial density $\rho_0(r)$, one computes the energy using (1.14), which is then used to solve (1.16), from which a new density is obtained using (1.13) which can be put into (1.16) to get another energy value. This self consistent-cycle proceeds until the difference between the energy of two successive cycles is below a user-defined threshold. The ground state density ρ_0 and energy E_0 can be taken as those of the last iteration.

In general, the method requires a basis-set to be used to represent the wavefunctions of the virtual electrons. There are two popular sets of basis-set that are currently used: atom-centered gaussians and plane waves. In general, the gaussian basis-sets are most common in chemistry and when dealing with molecules (as they are inherently localized,

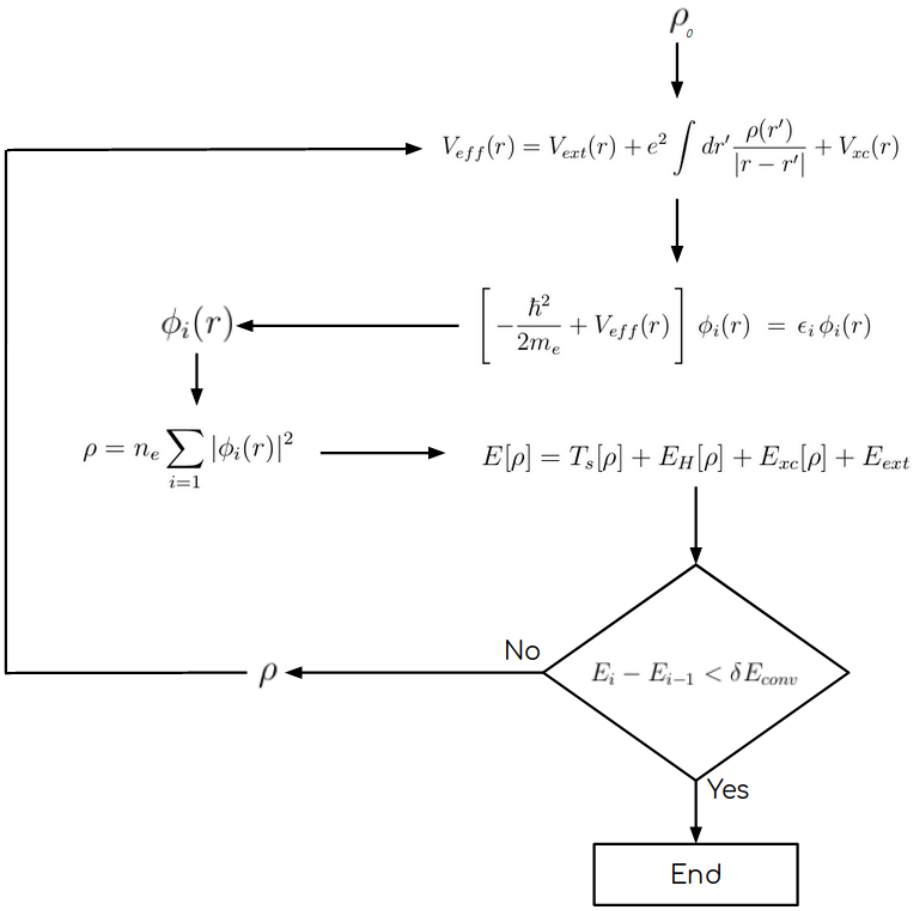


Figure 1.1: Illustration of the self consistent cycle of density functional theory

which suits the descriptions of orbitals) while in materials science, the plane wave basis-set is more commonly used. In this work we exclusively used plane wave basis-sets. Finally, the expansion of the basis-set is an important parameter for plane wave basis-sets, and it is generally up to the user to fix a cut-off on that expansion: in the case of plane waves, it determines how well localized behavior will be taken into account, but once again, it is in general fixed as a compromise between accuracy and computational cost.

1.1.2.2 Functionals

The one remaining issue is the expression of the exchange and correlation function $E_{xc}[\rho]$ and that of its functional derivative $V_{xv}[\rho]$. Although up to this point the resolution was general, and required no further approximation, as there is no known expression for those functionals, some will need to be used to go forward. In essence, there is a very large range of functionals type but the two most important kinds are the Local Density Approximation (LDA) and General Gradient Approximation (GGA).

In the LDA supposes that the density is roughly homogeneous and proposes that the LDA functional form of $E_{xc}[\rho]$ should depend only on the density:

$$E_{xc}^{LDA}[\rho] = \int d^3r \ \rho(r) e_{xc}^{HEG}(\rho(r)) \tag{1.17}$$

where e_{xc}^{HEG} is the exact Exchange and Correlation contribution per electron of the Homogeneous Electron Gas, which can be calculated through *ab initio* calculations such as Quantum Monte Carlo methods that exact but costly.

The main weakness of the method is in its name: it assumes that the density varies smoothly in space, which may be the case in some materials but is certainly not the case in molecules for example. It will however be a good model for systems where this assumptions holds like metals and semi-conductors. Another weakness of the method is that it does not describe well long range interactions as it is local.

The Generalized Gradient Approximation (GGA) is more complex functional: it uses as basis the LDA functional but adds a term that accounts for both the local density and its gradient in order to include the effects of variations of the density . In general the expression of the GGA functional is:

$$E_{xc}^{GGA} = E_{xc}^{LDA}[\rho] + \int d^3r \ e_{xc}^{GGA}(\rho(r), \nabla\rho(r)) \tag{1.18}$$

with e_{xc}^{GGA} the exchange-correlation energy per electron in the GGA approximation. There are several methods to construct this term leading to a host of different GGA functionals. In this work we mainly used GGA-PBE functionals [72], but we occasionally used GGA-BLYP[73, 74] to check some of our results.

This functional allows DFT to be used on a large spectrum of systems. However, it is always necessary to be aware of the fact that no functional is exact and of the inherent limitations of DFT.

In general, for example, DFT poorly accounts for long range interactions such as Van der Waals, and requires that additional (empirical) elements be added to properly describe them properly [75]. In the case of high pressure carbon dioxide, the effect of corrections of DFT functionals to account for Van der Waals interaction has been studied [76] and has proven to add a small, though non-negligible [18], changes in the overall properties of the various molecular and polymeric phases.

In the realm of *ab initio* methods, density functional theory is widely recognized as a very good compromise between computational cost and accurate description of the electron cloud. However, despite its relatively efficiency, the method is still expensive, and scales roughly with the cube of the number of electrons in the system.

1.1.2.3 Pseudopotentials

Pseudopotentials are an extra layer of approximation on *ab initio* calculations which relies on the chemical notion that the core-electrons do not contribute significantly to the chemical behavior of an atom. Although this *frozen core* approximation does not always hold, especially at high pressure, it is in most cases a good model and it allows for some important cost reduction for *ab initio* calculations by reducing the number of electrons that need to be taken into account in the system. This is all the more important given that those electrons tend to be very localized around their nuclei, and would therefore be particularly costly to describe using plane waves.

In general [2], pseudopotentials are built to reproduce the behavior of the core electrons using an smooth effective potential up to a cut-off radius r_c (the core radius). The potential does not oscillates in this region, so that it may be built using the minimum number of plane waves. The r_c parameter is by no means general and it is fixed and unique to each pseudopotential, as usual using a compromise between accuracy and computational cost.

There are two formules of pseudopotentials in the literature: Norm-conserving pseudopotentials [77], built so that the norm of the virtual wavefunctions are still normalized despite the use of the pseudopotential (hence their name); Ultrasoft pseudopotentials [78] on the other hand do not enforce the normalization of the wavefunction and are built in a way that allows for faster calculations, requiring a much lower cut-off on the expansion of the basis-set for plane waves. However, in general, the use of these pseudopotentials are more complex to implement into *ab initio codes*, which explains why they may not be used for all purposes despite their reduced computational cost.

1.1.2.4 Forces and pressure

Once an *ab initio* calculation has converged for a given nuclei configuration $\{R\}$, it is possible to compute the forces acting on atom i using the Hellman-Feynman theorem [79]:

$$F_i = -\frac{\partial E(\{R\})}{\partial R_i} = -\langle\psi|\frac{\partial H}{\partial R_i}|\psi\rangle \tag{1.19}$$

From those forces it is possible to perform a geometry optimization (or relaxation) of the atomic configuration by using a gradient descent method such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS)[80] or Conjugate-Gradient (CG) algorithm to move the atoms in such a way as to minimize the norm of the forces acting on the atoms and thus the energy. Equilibrium is reached whenever a criterion defined by the user is reached between two successive iterations of the algorithm. This is particularly useful when trying to find a stable configuration which corresponds to a stable crystal form, for example.

[2] again, more precision may be found elsewhere [52, 53].

In general, in condensed matter, those relaxations are performed at a given pressure. Pressure acts on the simulation box and forces this way, and also affecting the electronic density in the process. Following [81], one can compute the electronic component of pressure by first computing the stress tensor which is a 3x3 matrix defined as:

$$\sigma_{i,j} = -\frac{1}{V}\frac{\partial E}{\partial \epsilon_{i,j}} \tag{1.20}$$

where i and j are cartesian indices, $\epsilon_{i,j}$ is the strain on the simulation box, and V is its volume. The diagonal term corresponds to compression or dilatation of the box, while the off-diagonal term relates to its shape deformation.

The pressure is then computed as:

$$P = -\frac{1}{3}Tr(\sigma) \tag{1.21}$$

This expression does not account for contribution of the temperature (or the movements of atoms) to the pressure. An additional term is therefore necessary whenever one is studying systems using molecular dynamics (see 1.2).

1.1.3 Electron Localization Function

Although the electronic density of a molecular system may inform about the general electronic structure of the system, it does not describe very well the localization of the electrons. This may be an issue if one is studying chemical reactions as the localization of electrons gives important information about the electron cloud such as the presence of non-bonding doublets, covalent bonds and the type of bonds, lone pairs, etc...

In order to fix this issue, the electron localization function (ELF) have been proposed [82] with the explicit purpose of providing an accurate representation of electron localization in a system. The first proposition for the expression of the ELF was the following:

$$ELF(r) = \frac{1}{1 + \left(\frac{D_\sigma}{D_\sigma^0}\right)^2} \tag{1.22}$$

Where D_σ and D_σ^0 are the curvature of the electron pair density for electrons of identical spin σ for respectively the target system and the homogeneous gas (of same density) respectively. This approach results in values between 0, which corresponds to region between two electronic shells (due to the Pauli principle), while values close to 1 correspond to areas where, if one electron were to be there, no other electron of same spin may be found, which is caracteristic of lone pairs and bonding pairs. A value of 0.5 corresponds to a localization identical to the one of the homogeneous gas.

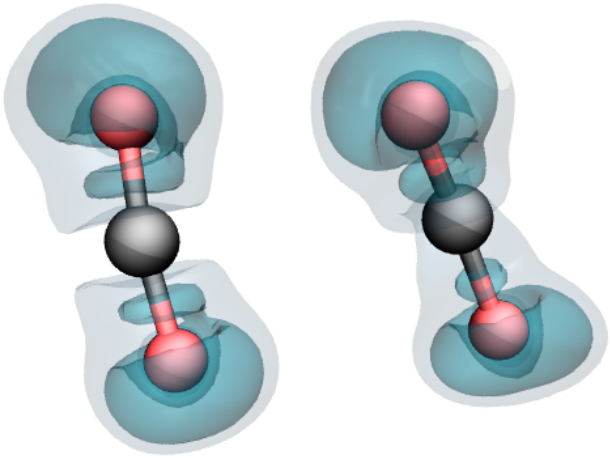


Figure 1.2: Isovalues surfaces of the Electron Localization Function of two carbon dioxide molecules. Two isolvalues are pictures here: 0.6 and 0.8

The issue with this definition is that the pair density and its curvature are not well defined within density functional theory. In order to adapt this method for DFT [56], it was proposed to change the interpretation of the ELF by taking a different expression for D [83]:

$$D = \sum_{i=1}^{n_e} \frac{1}{2} (\nabla \phi_i)^2 - \frac{1}{8} \frac{(\nabla \rho(r))^2}{\rho(r)} \tag{1.23}$$

Where ϕ_i are the Kohn-Sham wavefunctions for the virtual non-interacting electrons when using density functional theory, ρ is the electronic density and n_e the number of electrons in the system. D_0 is defined just as the same expression but in the case of the homogeneus electrong gas, as before.

The first term on the right hand side is the kinetic energy of the electrons while the term on the right is known as the Weizsacker kinetic energy density [84]. This term is exactly equal to the density of the systems if it were made of bosons, which can occupy the same quantum state while being in the same point in space. This implies that in places where the electronic localization will be high, this term will be roughly equal to the kinetic energy, and the resulting value of D will be 0, resulting in a value of 1 for the ELF. The ELF therefore measures of the excess kinetic energy density due to the effect of the Pauli principle.

As one can see on figure 1.2, the electron localization function has high values in the middle of a covalent bond, but it also shows high electronic localization around the oxygen atoms in shapes that can be associated with to the two non-bonding doublets typical of oxygen atoms. It therefore gives a chemically intuitive picture of the electronic cloud whuc can provides precious insight to understand the transformation occurring in atomistic simulations.

The ELF has been used in many different context to describe the electronic cloud: in order to give information on the various stages of bonding in small molecules such as H_2 and N_2 [85], or the polymerization of carbon dioxide [47, 85] but also providing insight allowing the characterization of the hydrogen [86, 87] and covalent bonds [88, 89, 90].

1.1.4 Force fields

Although *ab initio* calculations are very precise, they also tend to be computationally expensive, especially for large systems. It is for example not feasible to use *ab initio* calculations to simulate even small proteins in water. In order to simulate those systems, an alternative approach is necessary: that is were force fields come in. The aim of force fields is to provide simple functionals forms representing the interactions between atoms in such a way that it reproduces the results from *ab initio* calculations or experiments. Examples of such expression for the potential include the celebrated Lennard Jones potential [91], which can be used to represent long range interactions such as Van der Waals:

$$V(r) = 4\,\epsilon \left(\left(\frac{r_0}{r} \right)^{12} - \left(\frac{r_0}{r} \right)^{6} \right) \tag{1.24}$$

z where r_0 and ϵ are parameters that can be fixed by the user to reproduce the desired interactions between atoms.

This kind of approach can yield good results when the force field is constructed properly but suffers from a lack of transferability and flexibility to adapt to changing behavior, for example high pressure polymerization in the case of CO_2. Constraints on the angles or lengths in molecules may be added to the force field, in order to freeze degrees of freedom that are thought not to come into play in the transformations one is seeking to observe for example. In this study we used such a force field for the study of the transitions between the various molecular phases of carbon dioxide[92]: the force field is composed of only a Coulomb and Van der Waals contributions and with a restriction on the length and linearity of the CO_2 molecules as we are more interested in the weak interactions between molecules than in the internal modifications of carbon dioxide molecule itself.

Another approach is to use polynomial expansions methods in order to fit as much as possible the target potential energy surface. This kind of approaches has been shown to work well even for very complex system[93, 94], but they require a very large dataset of *ab initio* calculations (and the more precise the better) in order to give satisfying results.

Finally, with the avent of neural networks in the recents years, several propositions were made to train such networks to reproduce results from *ab initio* calculations and therefore act as force fields. This approach attenuates the need to find a good analytical form to represent the interaction by using the universal approximators[95] nature of the neural networks. Those approaches showed interesting results in their attempt to reproduce the PES of various systems, showing promise for future application[95, 96]. Meanwhile, some approaches successfully used machine learning techniques to bypass

entirely the Kohn-Sham equation[97]. Obviously, both of those methods require large amount of data to yield efficient results.

Overall, all force fields, regardless of their construction methods, are not transferable and correctly reproduce only behaviors on which they were trained. They also require large datasets of *ab initio* calculations to be parametrized. However they prove necessary when one wants to simulate systems on scales (spatial or temporal) that are beyond the reach of *ab initio* calculations.

Finally, it should be emphasize that they are very much *garbage in, garbage out*, and will exhibit the same defaults than that of the simulation on which they were trained: even the most precise force field trained on DFT data will still have the issues of DFT.

1.2 Molecular Dynamics

1.2.1 The general principle

In this section we present molecular dynamics which allows the calculations of thermodynamical equilibrium properties of a system using time averages of those properties. Indeed, assuming an observable A, we can write the time average A as:

$$\langle A \rangle_T = \frac{1}{T} \int_0^T A(t)\ dt \tag{1.25}$$

While the average value of a thermodynamical can be computed, noting q the coordinates of the system and p its momenta, as :

$$\langle A \rangle = \int \rho(\mathbf{p}, \mathbf{q})\ A(\mathbf{p}, \mathbf{q})\ d\mathbf{p}\ d\mathbf{q} \tag{1.26}$$

where $\rho(\mathbf{p}, \mathbf{q})$ is a probability density associated with the position and momenta. If the dynamics is *ergodic*, the two approaches are equivalent. In other words, it is equivalent to sample phase space to get the average value of an observable and to use molecular dynamics to evaluate over long periods of time the average value of the same observable. As the latter is generally easier to carry out, it may be preferred in some cases.

In order to compute this time evolution of the system, one simply needs to compute the evolution of the nuclei in time, using the forces from either a force field or *ab initio* calculations and by treating the nuclei as classical particles, calculate their evolution using classical mechanics. In the cases where *ab initio* calculations are used, this methods is called *ab initio* molecular dynamics (AIMD) while if force fields are used, it is designated as classical molecular dynamics (CMD)[3]

[3]This approach holds as long as the quantum nature of the nuclei can safely be neglected. In cases where it can't (low temperature systems with light atoms), one needs to resort to more involved methods such as Quantum Thermal Bath or Path Integral Molecular Dynamics.

In order to integrate the behavior of atoms using molecular dynamics, one starts from an initial configuration of nuclei $R(t)$ and their velocities $\dot{R}$ and mass M, and applies Newton's second law of classical mechanics:

$$M_i \frac{d^2 R_i(t)}{dt^2} = F_i(t) \tag{1.27}$$

for each of the nuclei i of the system. On then chooses a small time increment δt and apply the Taylor expansion to $R(t)$ to obtain the configuration at the next infinitesimal step of the simulation $R(t + dt)$:

$$R_i(t + \delta t) = R_i(t) + \dot{R}_i\, dt + \frac{\ddot{R}_i}{2}\, dt^2 + O(\delta t^3) = R_i(t) + \dot{R}\, dt + \frac{1}{2}\, \frac{F_i(t)}{M_i}\, dt^2 + O(\delta t^3) \tag{1.28}$$

From that point on, given that we have two successive step of simulation, we can use the Verlet algorithm [98] to solve the movements of the atoms for the rest of the simulation time:

$$R_i(t + \delta\, t) = 2R_i(t) - R_i(t - \delta t) + \frac{F_i}{M_i}\delta t^2 + O(\delta t^4) \tag{1.29}$$

$$\dot{R}_i(t + dt) = \frac{R_i(t + \delta t) - R_i(t - \delta t)}{2} + O(\delta t^2) \tag{1.30}$$

The important parameter here is the timestep δt, that should not be chosen too small so as to limit the numerical errors, but also not too large that the simulation loses its accuracy. In general, the timestep is taken between 0.5 femtosecond (fs) to 2 fs, depending on the constituents of the system. The initial velocities of the nuclei are chosen following a Maxwell-Boltzman distribution centered around the target temperature for the system.

On a technical note, if one uses a force fields that imposes constraints on the bonds and/or angles of molecules, the integration algorithm must be modified to dynamically maintain those constraint. In this work, we used the LINCS [99] algorithm as implemented in the GROMACS software [100] in order to do so. More information may be found on how those algorithm either in the GROMACS manual or in [101].

1.2.2 Temperature and Pressure

The "instantaneous" temperature can be computed from the results of a molecular dynamics simulations, using the equipartition theorem as:

$$T(t) = \frac{1}{3\, k_b\, N} \sum_{i=1}^{N} m_i \dot{R_i}(t)^2 \tag{1.31}$$

Where N is the number of nuclei, $\dot{R}_i$ and m_i the velocities and mass of the i^{th} nuclei respectively. However, the actual termperature is computed through time averages of this "instantaneous" temperature:

$$\langle T \rangle = \frac{1}{N\,\delta t} \sum_{i=0}^{N} T(i\,\delta t) \tag{1.32}$$

Depending on the statistical ensemble one wants to simulate, it might be desirable to use a thermostat to actively maintain the temperature at a target value. There are several possible alternatives, some using the rescaling of velocities in order to generate a distribution whose average is the target temperature such as the Berendsen [102] and velocity rescaling temperature algorithm [103] implemented in GROMACS [100] while the Nose-Hoover algorithm [104, 105, 106] introduces fictious masses in order to generate the proper velocity distribution.

In order to compute the pressure, one uses the virial theorem, and we find, for classical MD:

$$P = \frac{N\,k_b\,T}{V} + \frac{1}{3\,V} \sum_{i=1}^{N} \mathbf{F_i} \cdot \mathbf{R_i} \tag{1.33}$$

Where $F_{i,j}$ is the forces that acts on atom i from j and $R_{i,j}$ is the vector from i to j. The first term is the dynamical contribution of the pressure, while the second is the equivalent to the one displayed in (1.21), related to the stress of the simulation cell.

In the same way as for the temperature, it is possible to maintain the pressure close to a target value during a molecular dynamics simulation using barostat algorithm [107, 108] however, this implies to accept that the volume of the box evolves in time, which has been used as a good way to find new crystal structures.

When using a thermostat and or a barostat, it is a good practice to use a short simulation (a few ps in *ab initio* molecular dynamics, a few ns in classical molecular dynamics) to let the system equilibrate and reach the target temperature and/or pressure before actually starting a production simulation where results can be obtained.

Chapter 2

Topological based descriptor of the structure for physical systems

Introduction

In this chapter we introduce variables that aims at describing a system based on its topology, which we will use extensively in this thesis both to classify structures based on their similarity and in the context of enhanced sampling methods to enable the exploration of the energy landscape as described by those variables.

There are two different types of descriptors:

- variables that describe the system as a whole, which we will refer to as *global descriptors*, aiming at describing systems as a whole. They are generally associated with the study of the transformations of matter but are also used when classifiying structures or materials. Examples of such descriptors include the adjacency matrix-based variables such as the Social PeRmutation INvarianT (SPRINT)[55], Permutation Invariant Vectors (PIV)[54] or Coulomb Matrix[109].
- descriptor that capture the topology of individual subsets of systems such as atoms in a molecular systems for example. Those variables are referred to as *local descriptors*. Examples include the Symmetry functions used by Behler and Parrinello[95] or the Smooth Overlap of Atomic Positions (SOAP)[110].

2.1 Global variables

In this section we describe collective variables that depict the whole system at once. Global descriptors are used to quantify the distances, according to some metric, between two systems and to measure the amount of transformation required to go from one structure to the other. Applications of those variables include use as collective variable for enhanced sampling methods, classification of structures (crystals, molecules) and machine learning schemes.

There are two main symmetries that those descriptors must enforce : the invariance by translation of the whole system and the invariance by permutation of atoms of the same type (one can add in the case of molecules at least, invariance by rotation of the whole system). The most complicated symmetry to implement in practise is the permutation invariance as it implies the loss of some local information about the system.

All of the methods described below make use of a distance matrix, M, which is a matrix where each row and column correspond to a specific atom, and where each component corresponds to the distance between the row atom and column one : $M_{i,j} = d_{i,j}$ (with $d_{i,j}$ the distance between atom and j).

2.1.1 Social PeRmutation INvarianT coordinates

The Social PeRmutation INvarianT coordinates (SPRINT)[55] make use of graph theory[111] to describe the system by making an analogy between the network of interaction in a given structure and an adjacency matrix A. In graph theory, the coefficient of the adjacency matrix gives the connection between edges of a graph: if $A_{i,j} = 1$, it means that edges i and j are connected while they are not if $A_{i,j} = 0$. When continuous values in between 0 and 1 are allowed, they gives an indication on the "strength" of the connection between edges.

In order to build the adjacency matrix of a structure, one starts from the the distance matrix M and then applies a sigmoid function σ (also called switching function in the litterature) to all the distances. If several chemical species are present, one can use a different σ for each pair or specie. The parametrization of the switching function is left at the discretion of the user, as it allows one to choose what are the ranges of interactomic distances of interest depending on the situation. A common choice however is to associate a value of 1 or 0.9 to the first shell of atoms and a value around 0.3 for the second shell. Once the switching functions have been applied, one has the adjacency matrix A.

In general, the shape of the sigmoid function used for the normalization of the distances is the following:

$$\sigma(d_{i,j}) = \frac{1 - (\frac{d_{i,j}}{d_0})^n}{1 - (\frac{d_{i,j}}{d_0})^m} \tag{2.1}$$

where d_0 roughly corresponds to the bonding threshold and the values of n and m are used to control the smoothness of the sigmoid function. For obvious reasons, the user may define different sigmoids for each of pair of types of atoms in the simulation to account for the various kinds of interactions that can take place.

The only requirement on the smoothness on the switching function for the construction of the SPRINT collective variable is that the network associated with the matrix A needs to be connected: there should be a path from any point of the graph to any other following the connections (even weak) between the edges. This is equivalent to say that the network should not be composed of two disconnected networks. This requirement is

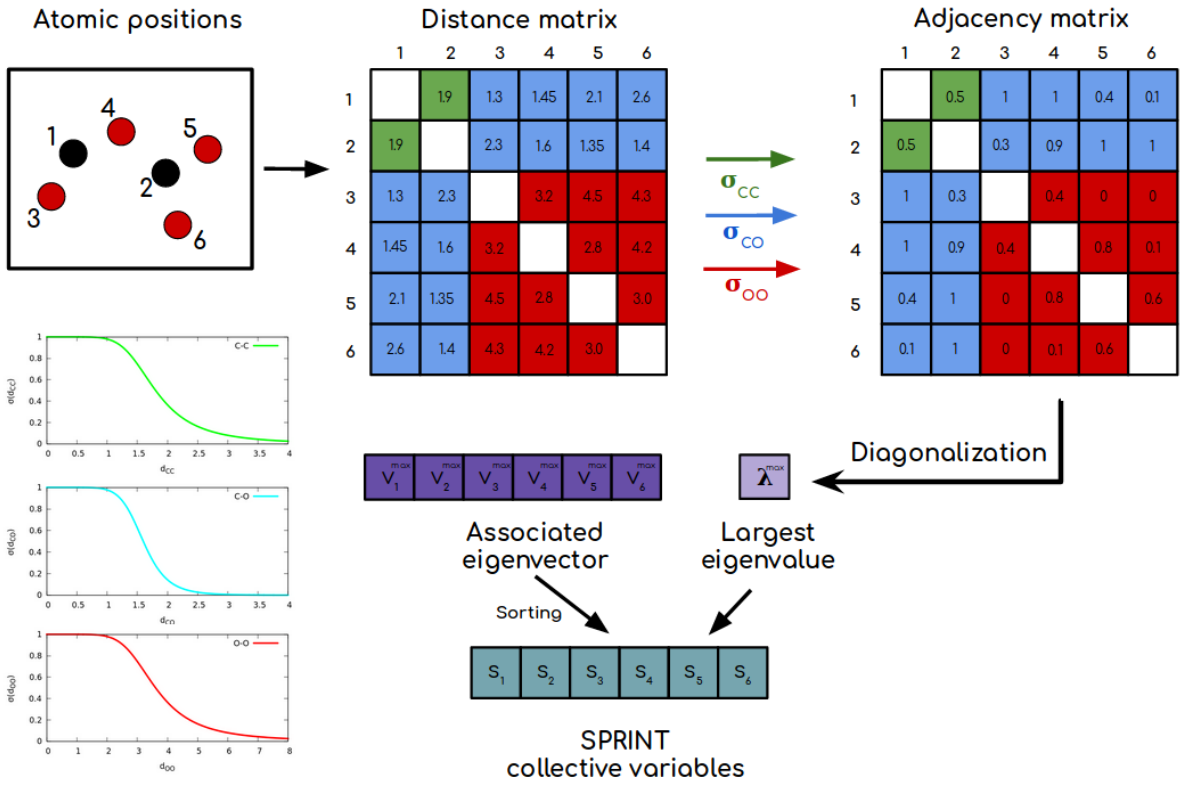


Figure 2.1: Illustration of the construction of the SPRINT collective variables

necessary for the Perron-Frobenius theorem to apply.

When this condition is met, the matrix is symmetric, non-negative and its associated graph being connected, the Perron-Frobenius theorem applies, implying the following properties about the largest modulus eigenvalue λ^{max} and the associated eigenvector $\nu^{\mathbf{max}}$:

- λ^{max} contains information about the network as a whole: it is comprised between the average and maximum coordination number in the system, and it increases with growing number of bonds in the system.
- ν_i^{max}, the ith component of $\nu^{\mathbf{max}}$, holds information about both the short and long range topology of atom i [1].

The SPRINT variables are then constructed by using λ^{max}, associated eigenvector components $\nu_i^{max,sorted}$ which are sorted from smaller to larger to enforce permutation symmetry:

$$S_i = \sqrt{N}\,\lambda^{max}\,\nu_i^{max,sorted} \tag{2.2}$$

[1] In particular, for any walk of length M on the graph, we have the following relation: $\nu_i^{max} = \frac{1}{(\lambda^{max})^M}\sum_j^N a_{ij}^M\,\nu_j^{max}$ where a_{ij}^M is the number of walks of length M connecting atoms i and j

where N is the number of atoms in the system, resulting in one collective variable per atom in the system combining both information about the connectivity of the system as a whole and the one of each atom. It is possible to compare two structures using SPRINT by simply computing the Euclian distance between two structures.

The combination of local and global information about each atom is sufficient to provide insights about its "social" network , through a more elegant way than for example, the coordination number.

The SPRINT coordinates have the interesting properties that each of its S_i is related to a specific kind of topological environment: if two S_i have the same values, it is highly likely that the associated atoms have the same local topology. Therefore, highly symmetrical systems or systems composed of similarly chemical units will have a large number of S_i degeneracies. On the opposite side, disordered systems will show highly different S_i values. This property makes it possible to extract some information about system using the time evolution of the S_i, even if the spanned space has a high dimension.

One should note, however, that the diagonalization implies both some part of loss of information and a large computation time, making those variables improper for very large systems. It also suffers from the fact that it does not account for any angle-related changes and is therefore not suited to study systems for which the transformations are mainly small changes in distances and angles with little bonding change. It would be for example problematic to use the SPRINT coordinates for the moleclar crystalline phases of carbon dioxide, where most of the transformation occur through change in relative orientation of CO_2 molecules.

The SPRINT variables have mainly been used for two applications: classifications of structures (molecules or materials) and exploration of configuration spaces using enhanced sampling methods[55]. Notably they were used for both purposes in a this work in a project unrelated to the thesis matter: the exploration of the configuration space of small MoS_2 clusters (see 7).

An implementation of the SPRINT collective variable can be found in PLUMED[112] for enhanced sampling and configuration space exploration purposes and in the piv_clustering code [54] for clustering ones.

2.1.2 Permutation Invariant Vector

Although SPRINT coordinates are very effective in their own rights, they have important limitations, and are not necessarily suited to describe the transition of bulk materials as those of molecular crystal phases of CO_2. An alternative is the Permutation Invariant Vector (PIV) [54], which is a very general and robust collective variable.

In order to build the PIV (figure 2.2), one starts on the same premise as the SPRINT

coordinates with the adjacency matrix A: we also start from the distance matrix and use a switching function δ to normalize the distances between 0 and 1 and obtain a normalized matrix. Here, again, the switching function should be defined as very smooth, to account for the long range topology around each atom.

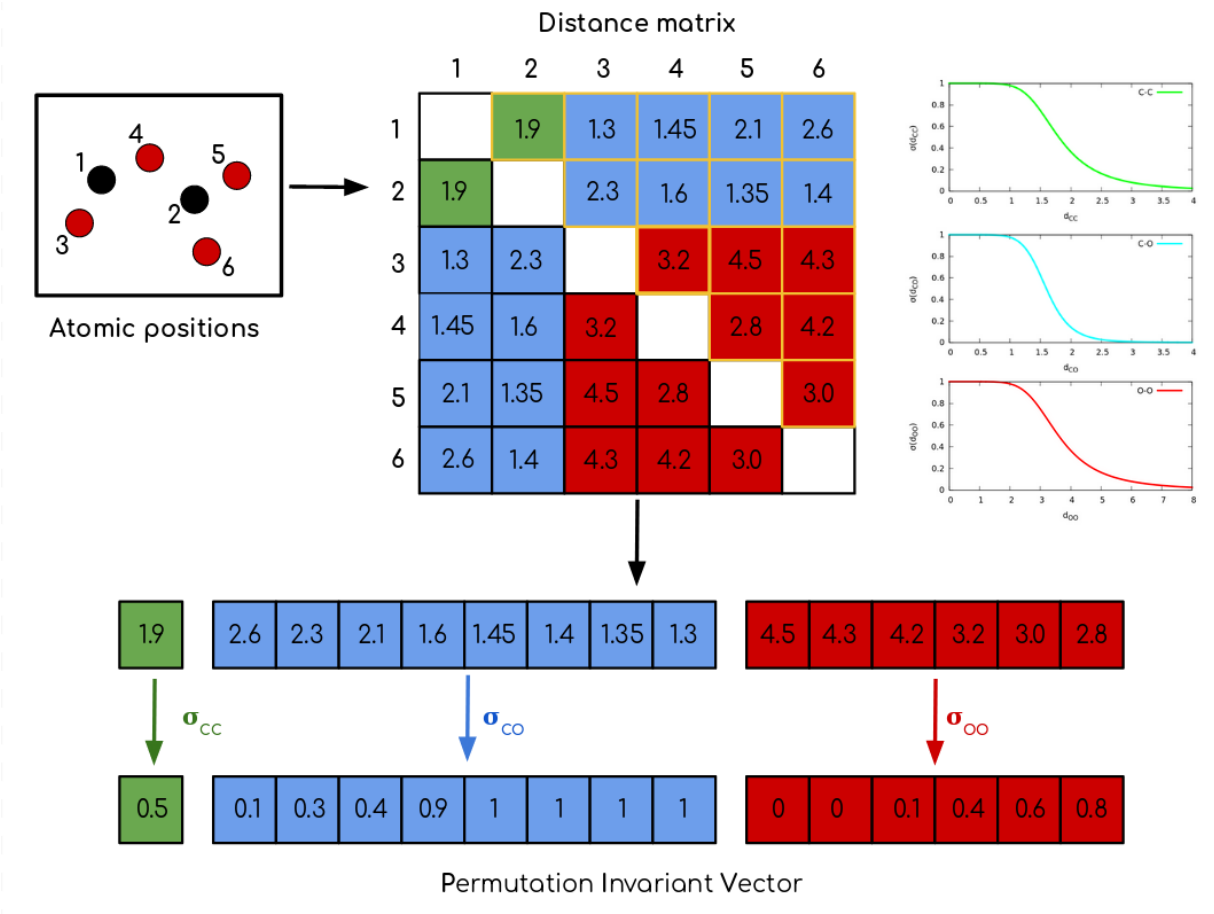


Figure 2.2: Illustration of the construction of the **P**ermutation **I**nvariant **V**ector

From the adjacency matrix A we extract the diagonal superior matrix (as the matrix is symmetric and the lower half is redundant) and separate it into blocks for each pairs of chemical species in the system. Each of the resulting blocks are then sorted and the set of blocks concatenated to build the PIV (figure 2.2). This resulting vector is very large: $\frac{N(N-1)}{2}$ elements in all with N the number of atoms in the cell. In effect, the elements of the PIV vector are written as:

$$v_{i,j}^{\alpha,\beta} = c_{\alpha,\beta}\ \sigma\left(\left(\frac{V}{V_0}\right)^{3/2} ||R_i^\alpha - R_j^\beta||\right) \tag{2.3}$$

Where α and β relates to the chemical species, i and j relate to the atom index with $i < j$, V and V_0 are the volume of the simulation box and a reference volume respec-

tively. R_i^α and R_i^β are the cartesian positions of atoms i and j of chemical specie α and β, respectively. σ is a sigmoid as used in the SPRINT coordinates and $c_{\alpha,\beta}$ is a scalar that allows one to scale the relevance of specific specie-specie interactions (for example, if the interaction between oxygen and carbon is not as important as the one between carbon atoms, one could set $c_{C,C} = 1.2$ and $c_{O,C} = c_{C,O} = 0.8$). The volume rescaling term $\left(\frac{V}{V_0}\right)^{3/2}$ is used in order to be able to compare systems that have largely different volumes.

As for SPRINT, the topological distance between structures or phases is computed by calculating the Euclidian distance between the PIV vectors associated with each structure.

In practice, computing PIV using this method becomes unfeasible for systems larger than a few hundred atoms,due to the potentially very large amount of sorting required on very large vectors. In applications where numerical efficiency is important, PIV is therefore computed slightly differently:

- The matrix M is divided into the type specific blocks as described above.
- The user chooses a precision on the value of the switching function. For each block a histogram with boxes of width corresponding to that precision are created: if the precision is set to 1000, the boxes will have a size of 0.001.
- The values of the $\sigma(d_{i,j})$ are used to fill the histograms for each types.
- The PIV is built with an integer vector built using the histograms of each blocks: the number of components for each integers is the number in the corresponding histogram box.
- Distance between PIV is computed using the Euclidian distance between the integer vectors of each structure, each element multiplied by the precision.

The main advantage of this method is that the sorting is done through the use of histogram and is therefore extremely fast and easily parallelizable, making efficient use of recent multi-core architectures of CPUs. However, even with this implementation, the vector size remains huge and the metric may not be appropriate for very large systems.

This metric contains a huge amount of information on the system's structure and, when the switching functions parameters are chosen appropriately, yields very little degeneracy, except the ones due to the limitation of using only distances as mentioned above. A proper choice of the parameters of the switching functions is important and can be tweaked by the user to include the range of distances relevant to the target transformations and if necessary ignore distances that are not relevant (like the intramolecular distances in a classical model where they are fixed). Once again, one may chose different sigmoid functions for each of the blocks, in order to account for different characteristic distances.

The main drawback of PIV is its size, making it relatively expensive to compute for large systems. Indeed, it is a very high dimension vector: $\frac{N(N+1)}{2}$ dimensions (with N

different atoms in the cell). In some cases the user may chose not to count types of atoms that are not relevant to the target transformations, which will speed the calculation by reducing the dimension of the PIV. This high dimensionality makes it also very difficult to visualize the PIV.

It should be noted that the sorting operation reduces the the generality of the metric as it is theoretically possible to construct pairs of structures that have an identical PIV[110] although they are different. Those structures are sufficiently very rare however and can be safely ignored in most practical uses.

This metric is rarely used in enhanced sampling by itself (due to its very high dimension) but more commonly within the Path Collective Variable (see below) to observe the transition between two or more structures. It was for example used with success to study the various phase transition in water ice under high pressure and navigate its phase diagram[113]. Interestingly, by projecting the PIV distances into a 2D plan it is often possible to get a topological map of the structures that bears a stricking ressemblance to the phase diagram[113].

The PIV metric is available in the PLUMED[112] plugin for enhanced methods and data analysis (through the driver), and the piv_clustering[54] for clustering.

2.1.3 Path Collective Variables

The Path Collective Variables (PathCV) are a special kind of descriptors best suited for enhanced sampling purposes: they were built in order to study specifically the transition from a given structure to another. They require the use of a function that is able, given two structures A and B, to return a topological distance, which we will note $D_{A,B}$. If one is studying the transition from a given state A to another state B and the simulation is currently in state $X(t)$, the two PathCV variables are defined as:

$$S = \frac{1e^{-\lambda D_{A,X(t)}} + 2e^{-\lambda D_{B,X(t)}}}{e^{-\lambda D_{A,X(t)}} + e^{-\lambda D_{B,X(t)}}} \tag{2.4}$$

$$Z = -\frac{1}{\lambda}\log\left(e^{-\lambda D_{A,X(t)}} + e^{-\lambda D_{B,X(t)}}\right) \tag{2.5}$$

S is related to the relative proximity of the structure X to structure A and B: if S is closer to 1, the phase is topologically closer to phase A while if it is closer to 2, it is closer to B. In practise, however, one uses the value of λ to place the reference structures in $S = 1.1$ and $S = 1.9$ so that the system can evolve around the position of A and B, to do so one sets $\lambda = \frac{2.3}{D_{A,B}}$

The Z variable measures the cumulative distance of structure $X(t)$ to both A and B: the larger the value of Z, the larger the distance of $X(t)$ to A and B. This variable allows the system to evolve away (or closer to) the target phases without specifically moving closer to any of the two phases, and is used to avoid forcing the system to move directly from one state to another, potentially exploring (using metadynamics (see 4.1.1) for example) easier path that are not direct in the space of the underlying collective variable space.

If one wants to include specific transitions through specific intermediary structures, it possible to use more than two structures. Denoting T_i the ith reference state, and given N the number of reference structures, the variables can then be written as:

$$S = \frac{\sum_{i=1}^{N} i\, e^{-\lambda D_{T_i,X(t)}}}{\sum_{i=1}^{N} e^{-\lambda D_{T_i,X(t)}}} \tag{2.6}$$

$$Z = -\frac{1}{\lambda} \log \left(\sum_{i=1}^{N} e^{-\lambda D_{T_i,X(t)}} \right) \tag{2.7}$$

Those variables are mainly used in either exploration or analysis of free energy landscape and have been used successfully in many different cases, ranging from biochemistry and pre-biotic chemistry [114, 115, 116] to phase transition in water ice [113] and are widely used within the PHYSIX team. In the present work they were used in the exploration of the phase transition between molecular crystals of carbon dioxide (see 6.2).

2.2 Local Descriptors

In this section we will describe a few variables that aim at describing the local environment around atoms, which can be useful for data analysis: either to determine the type of atomic behavior that exists in the system or how they evolve in time and in many other ways. This kind of descriptor is also used in order to relate local environments to the atomic energy through machine learning algorithms.

We note that it is generally possible to describe a whole of a system using a set of local descriptors, thus using a set of descriptors as global descriptor. The set should then be organized so that different atomic types are separated and permutation invariance is enforced. However, it is likely that the resulting descriptors may be overwhelmingly large and contain redundant information.

To begin with, one of the most simple way to describe the local chemical topology of an atom is the coordination number of a given atom for every type in the system, as it gives a first, coarse, description of its local environment. In practice, one can easily compute this for any atom i using:

$$C_k^i = \sum_{j \neq i} H(d_c - d_{i,j}) \tag{2.8}$$

with $H(x)$ is the Heavyside function, d_c a cut-off value that determines the maximum interatomic distance below which atoms are considered bonded, $d_{i,j}$ the distance between atom i and j, k relates to the type of atoms considered.

The set $C_i = \{C_1^i, ...C_N^i\}$ where N is the number of species in the system can be used as the first coarse descriptor of a local environment. In order to have more insight, several such sets with various cut-off values may be used to give additional information. It is also possible to use a smoother function instead of the Heavyside function: one can use a sigmoid function as in the SPRINT or PIV case or example.

This *first shell local descriptor* is relatively efficient to describe roughly the environement of atoms, it is not without its issues. The determination and use of cut-off distances can be limiting in some cases where the actual bonding depends on more than just the interatomic distance between atoms. This can be the case when the chemical state of each atoms has to be considered, or when thermal fluctuations cause the atoms to move around the cut-off distances frequently, leading to a spurious "flickering" of the bonding signal.

An alternative descriptor consists in using the sets of sorted distances to the N first nearest neighbors, N being a number chosen as the maximum number of atoms can be covalently bonded to (or direct interaction with) a central particle. Again, a more precise information can be obtained if those sets are computed separately for each chemical species present in the simulation.

This set $D_k = \{d_{i,1}, d_{i,2}, ..., d_{i,N}\}$ (with k refering to the target type as before) is also effecive in describing the local neighborhood and is less affected by flickering than the previous set, especially if N is chosen large enough.

In this work we used this metric when the determination of the coordination of carbon and oxygen atoms proved difficult, as basis for an unsupervised learning methods (see 3.2.3) to determine the coordination number of carbon atoms.

Finally, it is possible to regroup both descriptions to have information about the first two shells of atomic neighbors around a given atom i by computing, for each of the N first neighbor of type k, their own coordination number, andd then sorting the resulting set of second shell coordinations to enforce permutation invariance. An illustration of the resulting states is visible in figure 2.3.

For example, if one consider only carbon and oxygen bonding, and chose to care for the fist 5 neighbors of carbon atoms, the CO_2 molecule will be described by the set [1,1,0,0,0], as the carbon atom will have only two oxygen neighbor, each with one neighbor (the target carbon atom).

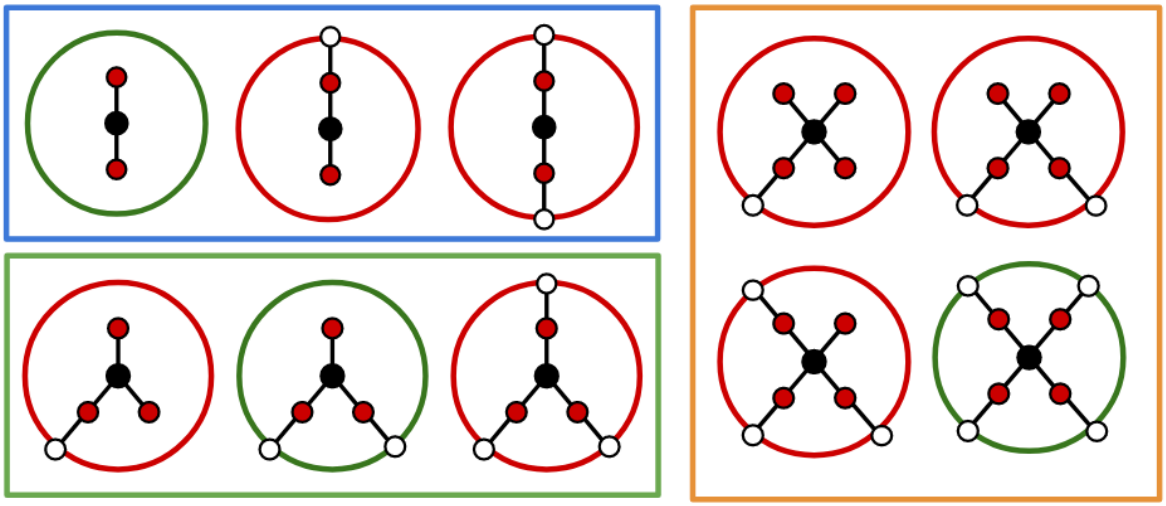


Figure 2.3: Example of carbon structures described with the second-shell local descriptor. Twofold coordinations in the blue rectangle, threefold coordinated carbon in the green rectangle and the orange rectangle contains fourfold coordinated carbon. The color of the circle around each structure relates to the presence (red) or absence (green) of charge in the structure. Carbon atoms are in black, oxygen in red, white atoms can be either carbon or oxygen.

Those *second-shell local descriptors* will be used to identify the various chemical states of carbon atoms in the Markov Model approach that we will use to gain insight on the complex chemistry of the highly relactive polymeric liquid.

Chapter 3

Statistical Learning Methods

3.1 Introduction

In this chapter we present various methods that aims at using statistical learning methods to extract information either about data sets through unsupervised learning or through time series with the Markov State Models.

We will use the unsupervised learning methods in practise in order to compare and classify structures in our projets related to search of structures and or exploration of free energy landscape. They will also allow us to analyze the local topology of atoms in the case of the liquid-liquid phase transition of carbon dioxide under extreme conditions.

Finally the Markov State Model will be used in order to analyze the time evolution of atomic states occuring in the polymeric liquid at high temperature in order to gain insight about the chemistry of said fluid.

3.2 Unsupervised Learning

Unsupervised learning is a class of methods that aims at learning information from a data set, without any external information about it. This kind of algorithm can be used for dimension reduction (with methods such as Principle Component Analysis (PCA) [117] or Singular Value Decomposition (SVD) [118]) or Classification (also called clustering) [119].

In this section we will focus on clustering, and more precisely on classification and tesselation of data sets. The aim of a classfication algorithm is to group together data points that are part of the same "structure" in space: given a set of spheres in space, an efficient classification algorithm should be able to group together all points related to the same circle. This is useful when one is trying to identify commonalities between data points, and to identify underlying similarities through statistical methdos.

A second application of clustering algorithms is tesselation - which is the process of dividing space into small elements much like tiles on a pavement. This is useful when one

is trying to reconstruct shapes out of points clouds, but also generally when dealing with large datasets, as using the cluster centers - that represents their local environement - of the tesselation may be easier than dealing with the whole set of points.

We present three different algorithm in this section: k-menoid/medoid [120, 121, 122], Daura's algorithm [123] and density peak clustering [62].

3.2.1 K-medoid algorithm

The K-medoid[120, 121, 122] is a very popular method to divide a data set into a user-defined number of groups which will optimize the partition using a cost-minimization process, where the cost is assigned to a given configuration of k clusters and corresponding data point affectation.

The main algorithm used in k-medoid clustering is the Partition Around Medoid (PAM) can be summed up as follow: one starts with a given configuration of k cluster centers, and the corresponding affectation of points and cost. Then, for each cluster center, the algorithm checks if replacing it with any other point that is not a cluster center lowers the cost. If it does, the new point becomes a cluster center in its place. This operation is repeated as long as the overall cost decreases.

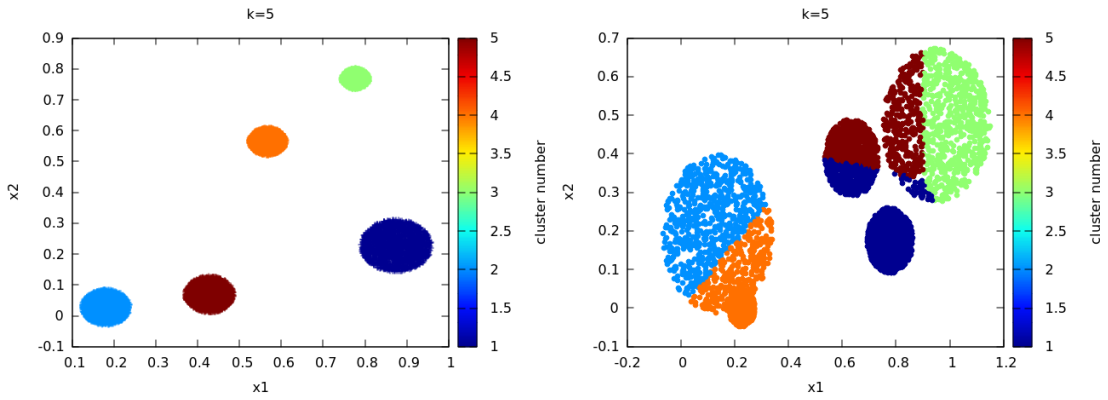


Figure 3.1: Examples of classification using k-menoid algorithm, in a case where the classification is efficient (left) and in a case where it fails (right)

The scaling cost of the method is relatively poor, at best scaling as N^2 with N the number of points in the data set. A faster alternative consists in simply checking the points within the original clusters, however, it severely reduces the ability to find the global minimum corresponding to the best partition. Indeed, points cannot be exchanged between clusters and therefore the algorithm explores a smaller search space.

There are four crucial points with this methods: the choice of a cost function, that of the assignment process for the points, the number of cluster center k and the original

seeding[120].

There are several choices for the cost function but the most commonly used is the sum of euclidian distances of all points to their respective cluster center. This has the advantage that it can be computed very efficiently, however it will inevitably favor a Voronoi-like partition of space.

The easiest method for assigning points to a cluster is to assign each point to the group of the closest cluster center, however this partition will also produce a hyper-spheric division of space around the cluster centers, which may or may not be the desired objective.

The number of cluster center is probable the most critical parameter of the algorithm. In order to choose it properly, it is a good practise, in the absence of prior information about the number of cluster one should require, the best way is to test several values and to analyze the variations of the classification with different values of k. Finally, the seeding of the original points is very important as it can decrease significantly the number of steps necessary for the convergence[120].

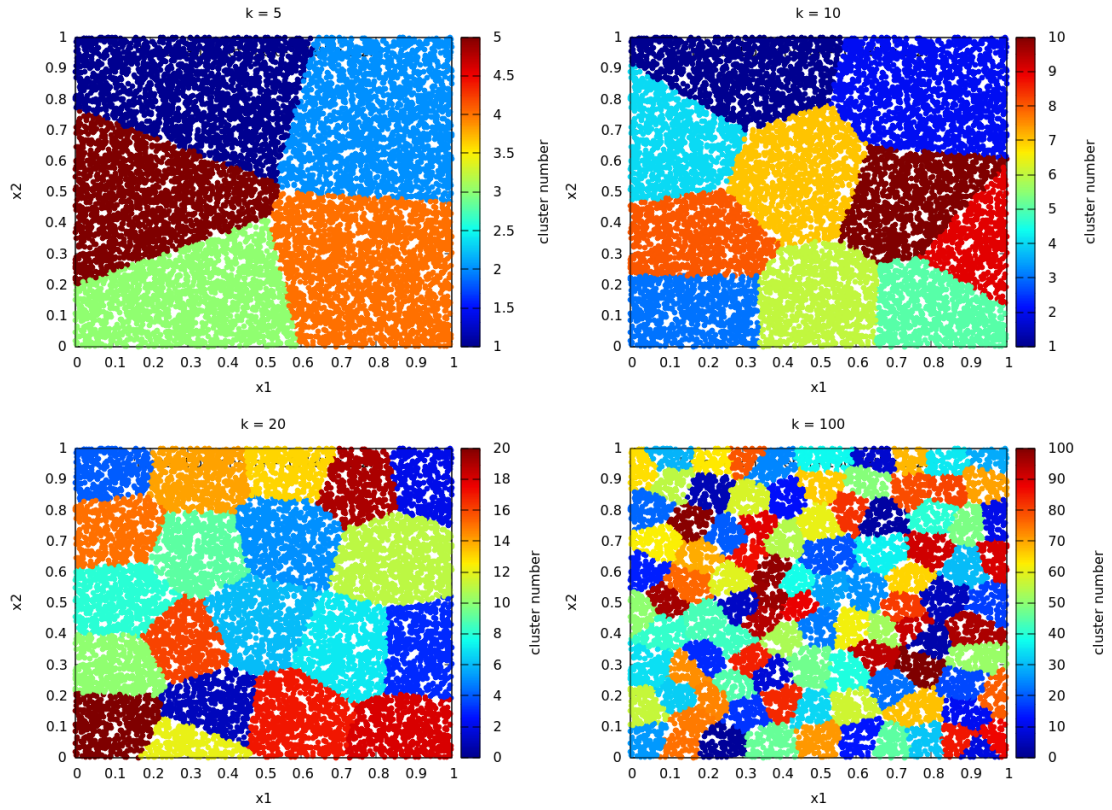


Figure 3.2: Example of space tesselation using k-medoid algorithm in 2D for k={5,10,20,100}

In general, it is not possible to determine that the final configuration corresponds to the optimal partitioning, therefore it is good practise to repeat the algorithm several

times whilst bookkeeping the best configuration and the associated cost, in order to see if a better configuration emerges.

In general, k-medoid is not a very effective classification tool (figure 3.1). It may work efficiently when clusters have similar sizes and shapes and share little overlap but in most cases it will not be able to actually recognise the underlying structures of the distribution of points. However it is an very useful tool to produce a tessellation of space (figure 3.2) with a given number of clusters as it will create clusters or roughly similar shapes and sizes in a very systematic way.

The algorithm is very popular and almost all machine learning libraries have an implementation of the method[124]. We also note that the piv_clustering[54] code provides an implementation for partition of structures using the PIV metric.

3.2.2 Daura's Clustering Algorithm

Daura's clustering algorithm [123] is an algorithm that aims at making a simple partition of a given point cloud, based on the density of the point cloud. It was used originally to identify several protein configuration from a molecular dynamics simulation. The algorithm requires a single parameter which is the cut-off distance d_c that determines whether or not two atoms are neighbors.

Daura's algorithm consists in identifying the point with the highest number of neighbors within a given radius d_c as cluster center, and assigning to its cluster all its neighbors. By repeating this step until all points are assigned to a specific cluster one end up with a classification of the point cloud.

In order to be efficient in terms of classification, this method requires that the data set is (hyper-)spheric shape. It is likely to fail in other, more complicated geometries, except if clusters are widely separated and d_c is chosen appropriately. It is possible to transform d_c into a vector $\mathbf{d_c}$ so that the radius describes an elliptic shape rather than a spherical one, but it remains that more exotic shapes (such as concentric rings) will likely cause the algorithm to fail for classification.

Much like k-medoid, this algorithm is not very efficient for partition of abstract datasets, although it may work in some cases (figure 3.3). It requires specific type of structures and a properly chosen vale of d_c so that the partition proposed by the system makes sense. Indeed, a larger value may be too large may hinge on nearby cluster, potentially affecting their cluster center, while a d_c value too low may results in the creation of very small clusters in between or at the edges of other clusters, especially when the density varies widely. Further, due to the importance of d_c, the target clusters should roughly be of the same size (or well separated), as if the variation on the size of the cluster is too wide, it becomes impossible to choose a d_c value that will allow one to reliably assign the points.

Another application of the method is the tessellation of space (figure **??**), much like

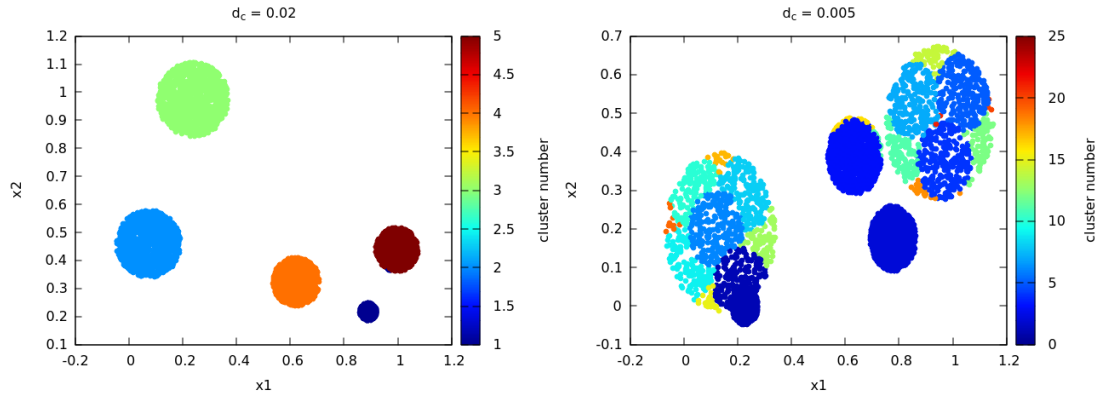


Figure 3.3: Examples of classification using Daura's algorithm, in a case where the classification is efficient (left) and in a case where it fails (right)

the k-menoid/medoid algorithm. The difference will be that the clusters will be spheric in this case and that one can chose precisely the maximum size of the unit volume/surface using the d_c parameter. If one wishes to make elements of roughly equal sizes, it is then necessary to use a relatively small d_c with regard to the dispersion of the point cloud in space in order. As before, it is recommended to test the clustering by variation of the value of d_c, and it is also important that the space is properly sampled to get an effective tessellation, as all points will be affected to the nearest cluster centers, which in turn is necessarily a data point. As Daura's algorithm is deterministic, it is also possible to check the quality of the sampling by evaluating the number and position of the cluster centers for increasing number of points, which is yet another possible application of this clustering algorithm.

Finally a last technical drawback of the method is that it tends to create clusters of very small sizes (even containing single points) with points being at the edges of space or between clusters. In general, it is relatively easy to find those clusters afterword, due to their very low sizes, and potentially to reassign the points to other clusters.

This algorithm is notably implemented in the code piv_clustering [54], much like k-means and was used notably in the course of this work to clusterize the trajectory from metadynamics simulations of MoS_2 nanoclusters (see 7) and in the context of the classification of structures in the project of improving the *ab initio* search of structure methods (see 8).

3.2.3 Density Peak Clustering

Density Peak Clustering (DPC) is a robust clustering algorithm proposed in 2012 [62] that aims at classifying data points based on their density, which is defined, like in Daura's algorithm, as the number of points within a cut-off distance.

The method relies on the expectation that density cluster centers should be character-

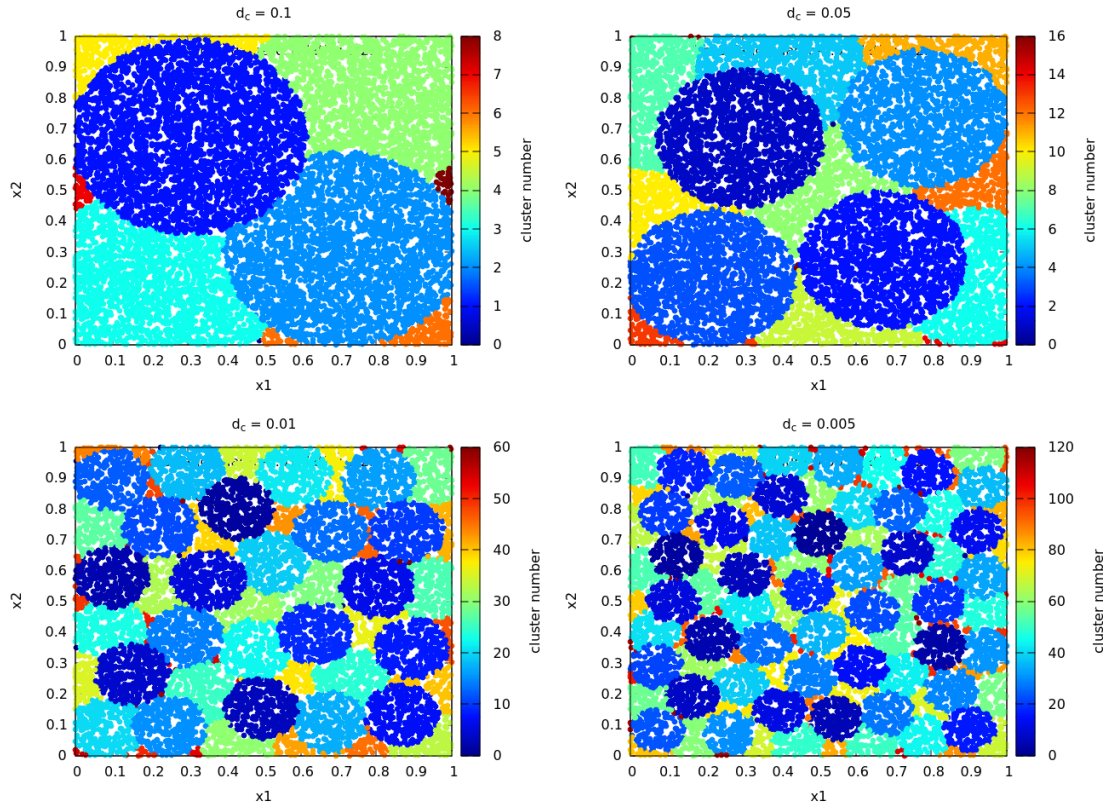


Figure 3.4: Example of space tesselation using Daura's algorithm in 2D for d_c={0.1,0.05,0.01,0.005}

ized by a high local density and by a large distance to the nearest points of higher density (as their immediate neighborhood should have a lower one).

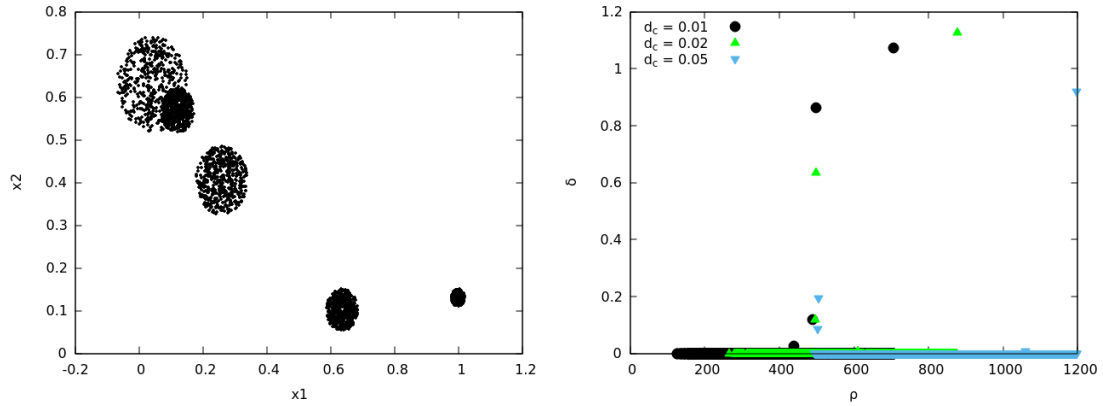


Figure 3.5: Example of a distribution of points (left) and the associated decision diagram for various d_c parameters (right)

In order to identify the cluster center the algorithm is made up of three successive steps. First, the local density (ρ) is computed for each point in the data-set, along with

the distance to its nearest higher density point (δ). Once this is done, one analyzes the plot of δ as a function of ρ , refered to as the *decision diagram* [62]. In this plot, the cluster centers should stand out as point with both high ρ and δ. Finally the points are affected iteratively to the cluster center of their nearest neighbor of higher density starting with the higher density points. This affectation process allows one to capture generally complex geometries[62].

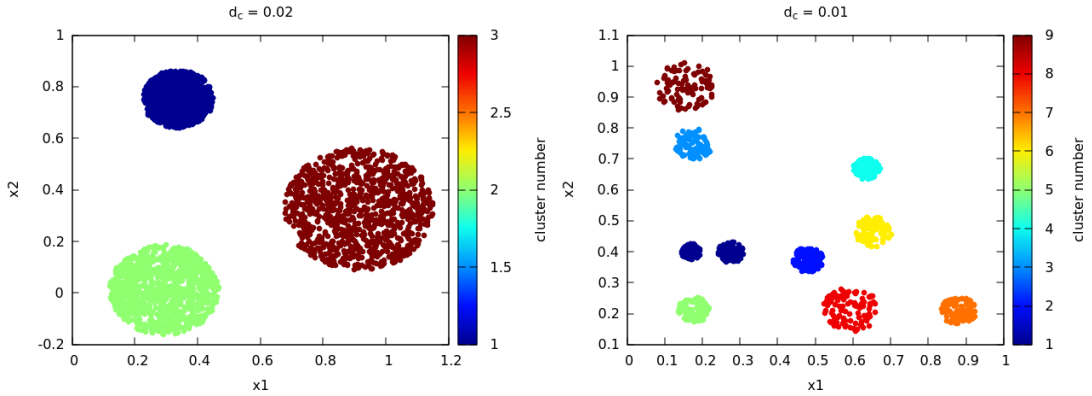


Figure 3.6: Examples of classification using Density Peak Clustering algorithm, in a case where the classification is efficient (left) and in a case where it fails (right)

There is therefore three main parameters for the algorithm: first d_c, then ρ and δ. A proper choice of those parameter is necessary as they determines the efficiency of the clustering and.

We will first focus on the computation of the value of the local density ρ, as δ depends on it. In the original paper [62], it was proposed to use the number of points within a given radius d_c of each points. This method has two main drawback: it is expensive to compute as it scales in N^2 with N the number of points in the system. Second, this makes the results of the method potentially highly dependant on the choice of d_c. This last part can be somehow alleviated by defining ρ for each by data point as:

$$\rho_i = \sum_{j \neq i}^{N} e^{-(\frac{d_{ij}}{d_c})^2} \tag{3.1}$$

This definition of ρ is smoother and captures contributions from several shells of neighbors around each point. However it does not removes the efficiency dependency on d_c. Prior statistical knowledge about the distributions of distances may be useful to determine an appropriate value of d_c

Once the value of ρ are known, calculating δ is straightforward. In some cases, a slight twist to the standard method may proves useful to help sort out the cluster centers. This twist consist in computing the distance to the nearest neighbor point of higher density

that is not within d_n of a higher density point, allowing the value of δ to be computed with regard to the nearest cluster center instead of the first point within the cluster that happens to have a higher density. Although this may not always be possible or straightforward, it increases the value of δ for cluster centers, especially in the case where two clusters are relatively close to each other. In general, it makes sense to use $d_n = d_c$, but using two different values may be better in some cases.

Finally, the user can identify the cluster centers on the decision diagram. The standard practice is to define minimum values of ρ and δ which serves as cut-off to delimit the cluster centers in the decision diagram. There is no general method in order to do so, therefore it is good practise to try several minimum values of ρ and δ to check the resulting classifications.

In general this algorithm proves robust and able to classify data sets forming many different shapes[62] (figure **??**), with relatively little human intervention. However it is not perfect and may fail in some cases. The algorithm proved popular and has been adapted to a large variety of applications[125, 126, 127], and many improvements of the algorithm have been proposed for various purposes[128, 129].

3.2.3.1 Finding local minima using DPC

The general spirit of DPC may also be used to find minimum in a sampled space by solely using the points that have been evaluated. This method that we believe is original, relies on the notion that there a local maximum (or minimum) of a function is characterized roughly by the same properties as ta local density cluster center: the only difference in this case is that the density is not computed through the number of neighbor but using the functional form of the function that one seeks the minimum or maximum of.

As far as we are aware, this approach has not been proposed before and although not foolproof, constitutes an interesting although modest contribution of the authors of this work to the litterature. This method was successfuly used in a project on the boosting of a search of (crystal) structure method (see 8) .

In order to find the minimum (or maximum) of a given function set using DPC the recipe we propose is as follow:

- 1 - Randomly sample the landscape and compute for each point its value, which will be used as the its density ρ of DPC.
- 2 - Compute the δ as one would normally do in DPC (with the nearest point of *lower* ρ if one is looking for a minimum) . It is also possible to use the same trick as mentioned in the DPC method to allow δ values to relate to the distance to the nearest point with lower ρ (respecitively hgiher point) that is not within a given radius of another point of lower ρ (respectively higher) density, in order to have an easier identification of minimum.

- 3 - Draw the the decision diagram, the local minima/maxima should stand out as outliers, much in the same way as in DPC with significantly lower ρ and high δ
- 4 - Add more points to the sampling and check that the positions of the minimum do not move significantly in the decision diagram, once the space start to be completely sampled, the points should converge to a specific position.

The advantage of the method is not necessarily in its efficiency as it may be easier and faster to make a large amount of steepest gradient descent in the landscape to identify local minima in many case. However it has the benefit of being a global method where the whole of the space is considered at once when assessing which points are minimum, allowing one to potentially eliminate shallow minimum that may exist around the global minimum. It also has the advantage to provide an convergence criterion for the exploration. Finally in cases where it is not possible to compute more in point than there already is in the landscape, it may be used to identify the minimum and get a picture of the landscape (although it is still highly dependant on a good sampling of the landscape to be effective).

Although the results we obtained with this method are very encouraging, the method is not foolproof: it may fail in cases where a golf hole-like minimum exists (that is a minimum with a very narrow well), and/or in cases where the overlap between wells is strong. In a general fashion, the method requires the ability to sample efficiently the whole of the space one wants to explore, which is not always practical.

3.3 Markov State Models

In this part we introduce the some basic elements of the Markov State Models (MSM), we refer the reader to the following reference [130, 131] for more detailed approach on the matter.

A Markov state model is a way to mathematically represent a system that is evolving randomly in a set of states without memory: the probability to move from any state i to another state j, $P_{i,j}$ does not depend on the history of states that was previously visited by the system. If we note the fact that the system is in state i at time t as $\mathrm{S}_i(t)$ we can write:

$$P(S_i(t)|S_j(t-dt)|S_k(t-2\,dt)\,|\,...\,|S_l(t-N\,dt)) = P(S_i(t)|S_j(t-dt)) \tag{3.2}$$

No matter what states i, j,k and l are.

A markov state model is defined by the following elements:

- A set of states $S = \{S_1 \,...\, S_N\}$ which the system will evolve in
- A transition matrix $mathbfT$ that holds the probability to go from any state i to any state j in the corresponding T_{ij}

- A set of prior $\pi = \{\pi_1, ... \pi_N\}$ that are the probability for the system to start in a given state i.

This type of systems were used successfully in to recover informations about the kinetics of proteins [60, 132]. In the context of this work, we will use them to analyze the kinetics of atomic states, in order to gain information about the chemistry of our systems, more specifically in the case of the polymeric liquid phase of carbon dioxide.

In practice we will be interested to analyze the evolution of atoms within a simulation cell, which we will observe through a sequence of states E, previously constructed using *local descriptors* (see 2.2) by assuming that this evolution can be described using a Markov State Model.

In order to be able to have use this description, one must first describe the transition matrix $T_{i,j}$. This can be done by first defining $\chi(S_1, S_2)$ functions that returns 1 if S_1 and S_2 are the same state and 0 otherwise. The expression of the elements of $\mathbf{T}$ is then [60, 132] :

$$T_{i,j}(\tau) = \frac{\sum_{i=1}^{M-\tau} \chi(E(t), S_i)\ \chi(E(t+\tau), S_j)}{\sum_{i=1}^{M-\tau} \chi(E(t), S_i)\ \chi(E(t), S_i)} \tag{3.3}$$

Which is equivalent to compute the fraction of states that at where in state i at time t and had transited to state j at $t + \tau$, over the whole set of possible transition.

From this point, one can check whether or not the system is markovian by checking that the following equation holds [60, 132]:

$$T_{i,j}(\tau) = \sum_{k=1}^{N} T_{i,k}(\tau/2)\, T_{k,j}(\tau/2) w \tag{3.4}$$

That is the probability for the system to evolve from state i to state j in a given time t is equal to the sum of probability for every "trajectory" that fulfill those boundary conditions.

If a system is indeed Markovian, one can use the Markov chain properties to compute interesting quantities related to the kinetics such as the lifetimes of the various states and the mean first passage time starting from any state i to state j.

Chapter 4

Free energy landscapes exploration

Introduction

In this section we will cover a group of methods that are used to explore the a free energy landscape associated with a system of a given chemical composition. In general free energy landscape are (very) high dimensional spaces that can be widely different depending on the system: liquids are expected to present a large number of shallow wells lying close to each other while the landscape relative to crystals are expected to present few very large wells corresponding to the most stable structures, with small local minimum corresponding to various metastable states closely related to them.

In general the dimension of the energy landscape is unknown but is expected to be large and which makes it impractical to sample efficiently using brute force sampling. Further, at least when dealing with condensed matter systems, the potentially large height of the free energy barriers makes it impractical for molecular dynamics to be used as an exploration method as the simulation time necessary to overcome those barrier is much too large with the practical simulation timescale currently available.

It is therefore necessary to use clever methods in order to be able to sample the free energy landscape. In order to do so, several methods have been established, although their basic methodology varies considerably depending on their objective. Here we will focus on two such applications:

- **Search of structures algorithms** where the goal is to identify as many of the stable configuration for a given configuration (in general at specific pressure conditions) in the energy landscape. Here, most of them methods will be based on geometric relaxation and on sampling the landscape at 0K.
- **Enhanced sampling methods** which aims at observing the transformation between different states, and potentially reconstruct the energy landscape related to the transformation. Here the methods will be used in tandem with molecular dynamics in order to either be able to explore or sample more efficiently the landscape. They will also involve the collective variables presented in (2.1).

4.1 Enhanced Sampling Methods

Introduction

Enhanced sampling methods generally aim at exploring the transition between the various local minimum rather than explore space itself. They are used in conjunction with molecular dynamics to help the simulation overcome free energy barrier that are much larger than the kinetic energy of the system. In those case, the transition from one state to the other is a rare event, and would be difficult to observe in molecular dynamics as it would require large simulation time.

Those methods make use of the global descriptors introduced in (2.1) to represent the free energy landscape. Indeed, as the hearth of methods that we describe here is either to push the system away from explored configuration (metadynamics[133]), or restrain the system around a given one (umbrella sampling) it is necessary to be able to represent the structure and more importantly, to compute distances between configurations.

An effective choice of collective variable is therefore necessary, not only in term of ability to differentiate the various configurations but also in terms of numerical efficiency: indeed, both types of methods that we describe here are notoriously expensive in terms of calculation time.

We will mainly focus here on the description of two methods: metadynamics and umbrella sampling, but the range of methods in this field is rather large and allow one to cover a wide range of applications. Metadynamics is mainly used in order to explore a landscape by progressively pushing the system out of the explored configurations. Umbrella sampling, on the other hand, is generally used to reconstruct the free energy landscape.

4.1.1 Metadynamics

Metadynamics is a technique that aims at accelerating the landscape exploration of molecular dynamics by addind fictious repulsive potentials V_b around explore structure, forcing the system out of explored configurations (which corresponds to free energy wells) as illustrated in figure 4.1.

In practise, small gaussians repulsive contributions are added each N_{bias} step of the molecular dynamics simulation to V_b, which in turns ends up filling up the free energy wells in which the system originally was, thus allowing it to progressively overcome even large energy barrier and explore other free energy wells.

In order to be able to define the explored configurations, one needs to make use of the collective variables defined in (2.1). Once a proper descriptor has been chosen, the gaussian bias contributions are then added progressively around the position of the system in collective variable space every t. Therefore the artificial potential felt by the system at time t will therefore be:

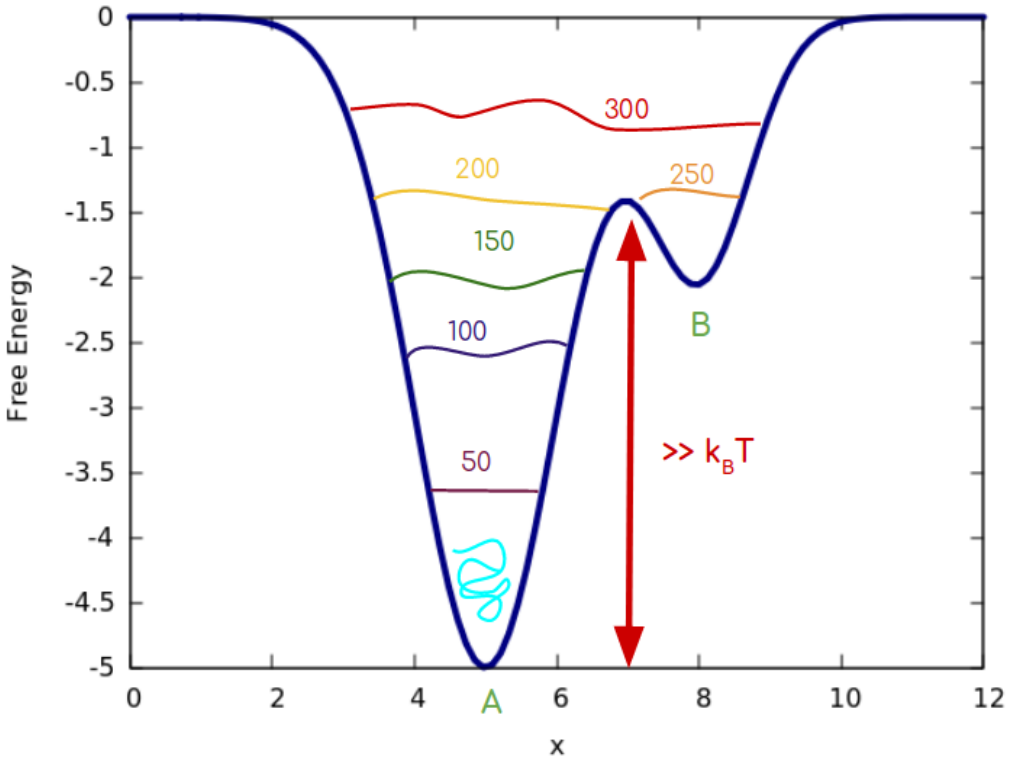


Figure 4.1: Illustration of the principle of metadynamics: the well in A is progressively filled up with fictitious potential until the system reaches well B. Numbers indicate the number of metasteps used to fill the free energy well up to the corresponding level.In cyan the trajectory of an unbiased molecular dynamics simulation for comparison.

$$V_b(S,t) = \sum_{t'=0}^{t} Ae^{\frac{(S(x(t))-S(x(t'))^2}{2\sigma}} \tag{4.1}$$

with V_b the fictitious potential, A the amplitude of the gaussian deposition, $S(x(t))$ some collective variable and σ the spread of the repulsive gaussian potential in the space spanned by the S collective variable.

It is worth noting that the quality of the exploration depends critically on the efficiency of the descriptor: a descriptor that with high dimension for example is not suitable as it would take too long for the potential bias to fill free energy wells in large dimensions.

Once the exploration of the system has finished, it is possible to reconstruct the free energy landscape by using the value of V_b much like a cast a mould [134, 135, 136, 137, 138]. For this method to provide a faithful portrayal of the FES, the metadynamics must converged - which correspond to a state where the algorithm has filled all the free energy wells and explore freely the FES. In practise, however, convergence does not necessarily

happens easily (it at all) and reconstructing the free energy landscape requires that the bias potential was put in sufficiently small increments so that the precision on the reconstruction is correct.

The user is free to choose the values of amplitude of the gaussians (A), the width of the gaussian in descriptor space (σ) and the regularity of the deposition of the bias potential. In general, larger values of A and σ will result in faster exploration but the precision on the sampling of the landscape will be lower, while reducing those values increase the required length of the simulation to observe the transitions but increase the precision for the reconstruction of the FES.

In the spirit of accurate representation of the landscape, well-tempered metadynamics[139] proposes, roughly, to use metadynamics with a progressively decreasing amplitude of gaussian fictitious potential, in order to help with the convergence of the system. This method requires some knowledge of the free energy landscape however to be use accurately.

Metadynamics has also been used as a purely exploratory method in order to identify new stable structures or phases, both in classical[55] and *ab initio* calculations [108]. The main drawback here is that metadynamics relies in essence on long molecular dynamics and the ability to explore unknown configuration space with this methods depends critically on the ability to compute long molecular dynamics calculations on the system at a reasonable cost.

Finally, it is important to note that metadynamics comes with a potentially large increase in computation time (and reduced simulation scaling) from the computation of the collective variables at every step: indeed, even if the potential is added only every N_{step}, the bias potential is effective on every step, and therefore (at least in theory) the collective variable must be computed at every step, potentially implying a large computation cost when it is computationally demanding (like PIV for example).Furthermore forces must also be computed in the collective variable space, which can be tricky and expensive. This can be somehow alleviated by only recomputing the collective variable every N steps, however, one needs to be careful that this value is not too large as to generate large errors.

Overall metadynamics is a powerful and somewhat popular algorithm due to its simplicity and efficiency. Its theoretical basis have been largely commented and studied along with its convergence properties[134, 135, 136, 137, 138]. It proved efficient on a number of different systems, targeting a wide range of transformation of matter [113, 55, 116] and in the case of both *ab initio* and classical molecular dynamics. In this work we used metadynamics to analyze the phase transition of molecular crystal phases of carbon dioxide (see 6.2) but also as a method to explore configuration space of small nanocluster (see 7).

An open-source implementation of metadynamics is available with the PLUMED plugin[112], that can be used with a large number of software, both for classical and *ab initio* molecular dynamics.

4.2 Umbrella sampling

Umbrella sampling[140] is a method designed to sample a specific area of the free energy landscape, in the space of a specified collective variable. The general idea is to define a potential bias of the form:

$$V(x) = \frac{k}{2}(S(x) - S_0)^2 \tag{4.2}$$

with a value of k relatively large (and chosen as a function of the temperature) that will force the system to stay with a given radius of the S_0. This potential is referred to as an umbrella (in reference to its shape).

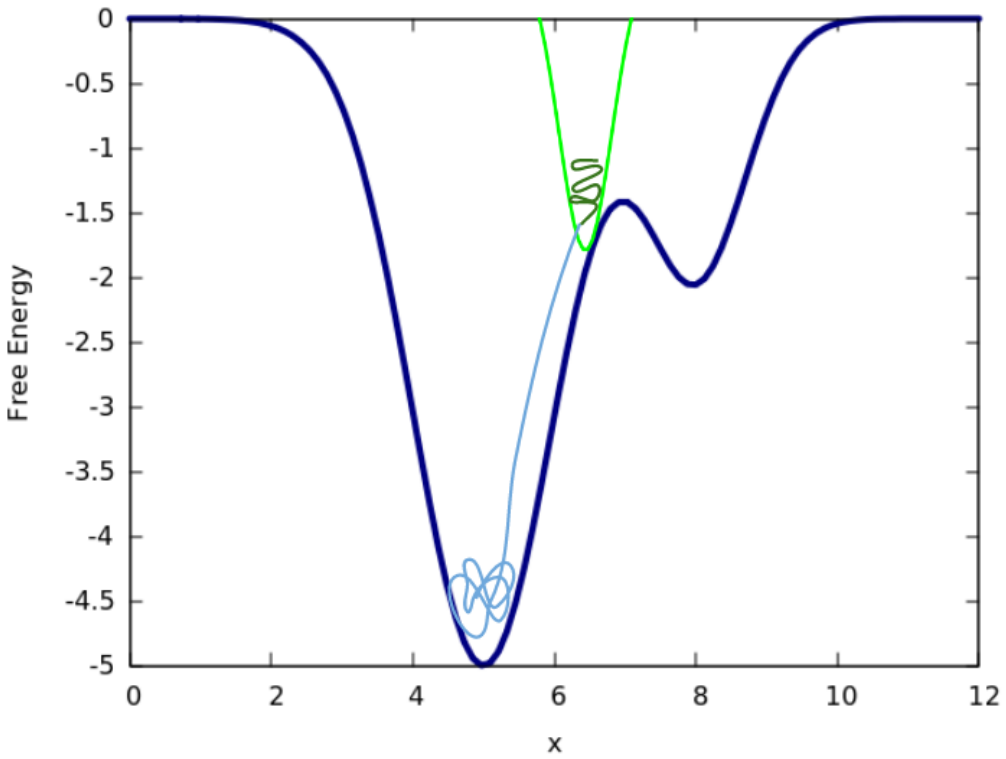


Figure 4.2: Illustration of the principle of umbrella sampling. In light green, a single umbrella, with the corresponding trajectory in dark green. The equivalent free molecular dynamics trajectory is shown in blue for comparison.

The resulting trajectory allows to have a good sampling of the small area where the system was allowed to evolve in. However, in order to choose properly the values of S_0 a first hand knowledge of the free energy landscape is necessary, which is why in general umbrella sampling is used in coordination with other enhanced sampling method, and often times metadynamics.

In most cases, one would therefore have a rough map of the free energy landscape of interest in the collective variable space, then use a tesselation of this space into small

fragments, and put an umbrella at the center of each fragment. Once all the umbrella have converged, the free energy landscape can be computed using the weighted histogram analysis method (WHAM)[141, 142, 143, 144, 145] .

Much like metadynamics, it is important to note that in general this procedure is more expensive than a simple molecular dynamics simulation, and in general, one of the downside of the method is that such a precise sampling of a free energy landscape, even only limited to two dimension is relatively expensive and should be done carefully.

4.3 Search of Structures

One of the main application of *ab initio* calculation is the use of its ability to predict the properties of material using very little input from the user. It is therefore no wonder that ever since they become computationally cheap enough to be run on local workstation, they were used to look for new and exotic materials, potentially with specific properties, in order to guide the experiments. In the high pressure field, they are extensively used both to look for new materials and to compare with present experimental data, in order to provide additional insights on the properties of structures.

Although nowadays relatively cheap, *ab initio* calculations do suffer from the shear dimensionality of configuration spaces, and their exploration tend to be computationnaly expensive, even for simple material, thus requiring specific algorithms in order to achieve this goal. Several methods were proposed over the years, some based on random sampling of the configuration space [61], others based on evolutionary algorithm [146, 147].

Here we present one such method that was devised for finding local minimum in such high dimensional spaces: the *Ab Initio* Random Searching of Structure. We note that, although we focus on search of structure at 0K, some exploration methods such as bassin-hopping[148, 149] and metadynamics [108, 150] can also be used to search for new stable structures at finite temperature.

4.3.1 *Ab Initio* Random Searching of Structures

The author is highly endebted to Adrien Mafety, a former PhD student in the PHYSIX team, for introducing him to the method and providing practical advice on how to use the associated software. The following presentation is inspired by his work[53].

Ab Initio Random Searching of Structure[61] (AIRSS) is in essence a very simple method. Starting from an initial guess of the target pressure and a given approximate volume of the crystal cell V_0, and the number of atoms of each atomic type, it consists in repeating a large number of time the following recipe:

- Build a random cell by randomizing cell parameters so that the total volume is within 5% of V_0. Boxes that are too skewed are discarded.

- Randomly put the atoms in the box (with a safety radius around each atom to avoid obviously problematic proximity between atoms).
- Run an *ab initio* relaxation of the system (which can be tuned to be complete or partial)

This procedure is stopped whenever the user is satisfied with the number of generated candidate structures. The obtained structures are then sorted by increasing energy, and the lowest lying structure are kept. In practise, one tends to keep the structure within a few eV of the lowest lying structure.

Duplicate structures are then eliminated, for example using their symmetry with the *findsym*[151] code and energy (or the PIV metric[55]). The remaining structures are then considered viable candidate structures, and one can proceed to further analysis.

The validity of the method rests on the fact that by taking enough random configurations and relaxing them, we are in effect sampling the configuration space and finding progressively all the minimum. The efficiency of the method relies on the assumption that all or at least the deepest minimum have an large enough attractive well around them that with a reasonable number of point, at least one ra dom structure will find itself in each of them. Although there is not clear theoretical evidence suggesting the validity of the method, it has proven successful in general and even when it fails to find to find some minimum, it is efficient to recover at least a couple of stable structures.

On the other side, the method has drawbacks: the cost of the method may fluctuate widely depending on shear luck and the individual cost of a single relaxation. As mentioned above the efficiency will largely depend on how efficient the random sampling is at finding all the various wells, but it also depends on the distance of the original random structure to the center of the well. Indeed, the cost of a single relaxation (and of that of an SCF) may vary depending on the initial configuration.

Furthermore, in the specific cases of narrow potential wells, the chances of finding said minimum randomly are slim. This point is amplified by the fact that the method is impeded by the scaling of DFT, which depends on the number of electrons. This means that the cost of the search will be prohibitive for system more than a few tens of atoms, even more so if the atoms have large Z. The method is also slightly at a disadvantage when dealing with a system where very competitive minimum have widely different density, as the method actively discourage this, as not imposing limits on the volume or imposing too loose ones would imply the exploration of a much too wide configuration space. It is also likely that system that tend to converge slowly (typically for magnetic materials), the cost of the method will also prove prohibitive due to the potential large computational cost of the relaxation.

In order to mitigate the cost of the method, several modifications where proposed. The most common one is to do a two-step search where the relaxation are first restricted to a small number of relaxation steps instead of targeting the fully relaxed structure. The

best structures are then kept, the duplicates discarded, and the remaining structures are relaxed to full equilibrium. This method diminishes overall cost by (potentially drastically) diminishing the cost of the most repeated operation. It is also noteworthy that by adjusting properly the V_0 value, large improvement on the cost of the search can be obtained. Finally the method can be tweaked if one is looking at molecular crystals in particular by randomizing not the position of atoms directly, but position and orientation of molecules in the simulation box instead.

The method has been successfully applied to a wide range of materials including carbon under TPa pressures [152], TPa phases of water ice[153] and solid hydrogen [154].

During the course of this work, a project within the PHYSIX team was dedicated to try and make improvement on the search of structure method (see 8) and AIRSS was also used briefly at the very beginning of this thesis on crystalline phases of carbon dioxide (see 6.1).

Part II

Results

Chapter 5

Liquid phase transformations

5.1 Context

We will report our work the transformation of carbon dioxide molecular liquid under the geological conditions of the lower mantle of Earth, more precisely we will focus on the range of pressure between 25 and 70GPa and 2000 and 3000 K. As mentionned above, understanding of the transformation and chemistry of carbon dioxide under those conditions is important for their implications on the properties of the mantle, notably for the formation of diamond [4, 5].

This region of the phase diagram coincide with the polymerization of carbon dioxide into polymeric phase at lower temperature and therefore we expect a similar behavior for the molecular fluid in this range. The low temperature molecular liquid was investigated mostly investigated in relation with the melting curve of the various molecular crystals of carbon dioxide [45, 45], although a mixed experiment and theory approach showed that it did transform continuously under compression, mimicking low temperature transformations of the crystal phase [6]. In this fluid both experiments and theory [6, 45, 45] suggest that molecules of CO_2 behave in much the same way as in the molecualr crystal phases : they interact through long range interactions (quadrupolar and Van der Waals) and the molecules are linear, with little bending of the molecule.

At higher temperature, running experiments proved more complicated and the few that managed to reach those extreme conditions indicated that CO_2 actually dissociated above 2000K [8, 9], although those results could have been skewed by interactions between the heating apparatus and the CO_2 sample, as experiments simulating Earth's lower mantle showed formation of carbon dioxide [155, 4, 5]. This hypothesis was recently reinforced by experiments showing that crystalline polymeric CO_2-V remains stable at 2000 K and 100 GPa [49]. In this context, where experiments are difficult to carry out, *ab initio* calculations are traditionnaly valued and two back-to-back theorerical investigations were performed on the high temperature behavior of the carbon dioxide liquid.

The first of those studies [7] showed through the use of AIMD calculations that the molecular fluid polymerizes into a polymeric fluid through a first-order liquid-liquid phase

transition [7]. This transition was characterized by a drop of twofold coordinated carbons, replaced by threefold coordinated and fourfold coordinated carbons. The first-order nature of the transition was evidenced by a plateau region in the P(V) relation[7]. This study also showed evidence that carbon dioxide did not phase separate at least up to 3000 K.

The second theoretical investigation focused on the analysis of the stability of phases IV, V and polymeric liquid near the geotherm[37]. Using AIMD and thermodynamical integration, it showed that CO_2-V was the most stable phase in much of the pressure-temperature range where the polymeric liquid phase was observed in the previous paper[7] and most importantly, in the range corresponding to geothermal conditions.

Although this study provides insight into the general behavior of geothermal carbon dioxide, we argue that energetic consideration do not take into consideration the kinetic or hysteresis effects. An important and relavant example is the theoretical prediction that CO_2-V is predicted by DFT to be the most stable phase for carbon dioxide as early as 20 GPa on the basis of 0 K relaxations, while it has only been observed over 40 GPa in practise. We therefore argue that analyzing the mechanisms of the transformations from the molecular liquid to the polymeric phases in order ot have the full picture of the behavior of carbon dioxide in geothermal conditions.

In this chapter, we will therefore study the transformations from the molecular to the polymeric liquid around the geothermal conditions using long AIMD simulation. Our objective is to provide additionnal information about the structural and dynamic aspects of the transformations occuring in these experimental conditions. In order to do so, we ran 100ps long *ab initio* molecular dynamics simulation in over 50 different simulations scattered on the whole range of pressure and temperature of interest. We will also closely analyze the behavior of the molecular liquid, as it was shown in [58] that carbon dioxide may form C_2O_4 dimers at slightly higher temperature (4000 K) than the ones we study, which may be indicative of of a different, more reactive, fluid than the one that has been studied until then.

5.2 Computational methods

We carried out *ab initio* molecular dynamics simulations, using the DFT level of theory with PBE functionals [72] and Martins-Troullier pseudopotentials [156]. We performed the simulations in the NVT ensemble, in a box using 32 molecules (96 atoms) and a Nose-Hoover algorithm to maintain temperature which frequency was set up at 2500cm^{-1}. Pressure was determined by the choice of a proper box volume. The cut-off for the wave-function was put at 120 Ry, after convergence tests in the most extreme condition (65 GPa and 3000 K).

The electronic convergence criterion was put at 10^{-5} Ry, and we chose a timestep of 1fs for all calculations (although some preliminary work was done using 0.5fs timestep).

The analysis of the simulation was done using a stride of 5fs. We let the simulation run for at least 10ps to equilibrate and ran production calculations for 100 ps.

In order to properly sample the pressure-temperature range of interest, we ran simulations at three different temperatures and 17 different volumes - each corresponding to a different pressure. All simulations started from the structures of the crystalline CO2-I phase. We used long thermalization of 10ps for equilibration, although longer thermalization times (up to 15ps) to insure proper thermalization in some specific cases.

5.3 Pressure-Volume relationship

As our calculations were done in the NVT ensemble, we computed the pressure for each simulation in order to choose proper volumes. As one the distinctive signs of the first order nature of the liquid-liquid phase transition in [7] was the existence of an abrupt change in the slope of the $P_T(V)$ relationship, we used this equation in order to have compare our results to the ones from this work.

The resulting $P(V)_T$ for each temperature of the study is shown in Figure 5.1 along with the one computed in [7] at 3000K. As in this study, we observe an abrupt change of slope, likely marking the transition between the two fluids. However while [7] locates it around 48 GPa, we find three different transition pressures: 48GPa at 2000K, 52GPa at 2500K and 56GPa at 3000K.

The difference between those results likely stems from different methodologies to initialize the simulation and/or from the length of the simulation. Indeed, in both study, the number of molecules, size of the box, and type of functionals (PBE) are identical but simulation times are much shorter in [7]: 2.5ps of thermalization followed by 10ps of simulation while this studies . It is also unknown how the simulations boxes were built in [7] which may also have an important impact the results.

5.4 Analysis of chemical bonds

In order to conduct a structural analysis of the molecules in the system it proved necessary to find an efficient definition of the bonds between atoms. The common practise is to use the first minimum of the partial pair correlation function as cut-off and this approach is reasonable in most cases [67]. However as we were dealing with a complicated system under extreme conditions we decided to investigate whether the high temperature-induced fluctuations would be sufficiently significant, and therefore test the validity of the use of cut-offs.

In order to illustrate the issue linked with high temperqture, we show in Figures 5.2 the distribution of distances of the first four oxygen to carbon atoms. We chose to compute those distribution on the most extreme conditions (65G GPa 3000 K) in order to

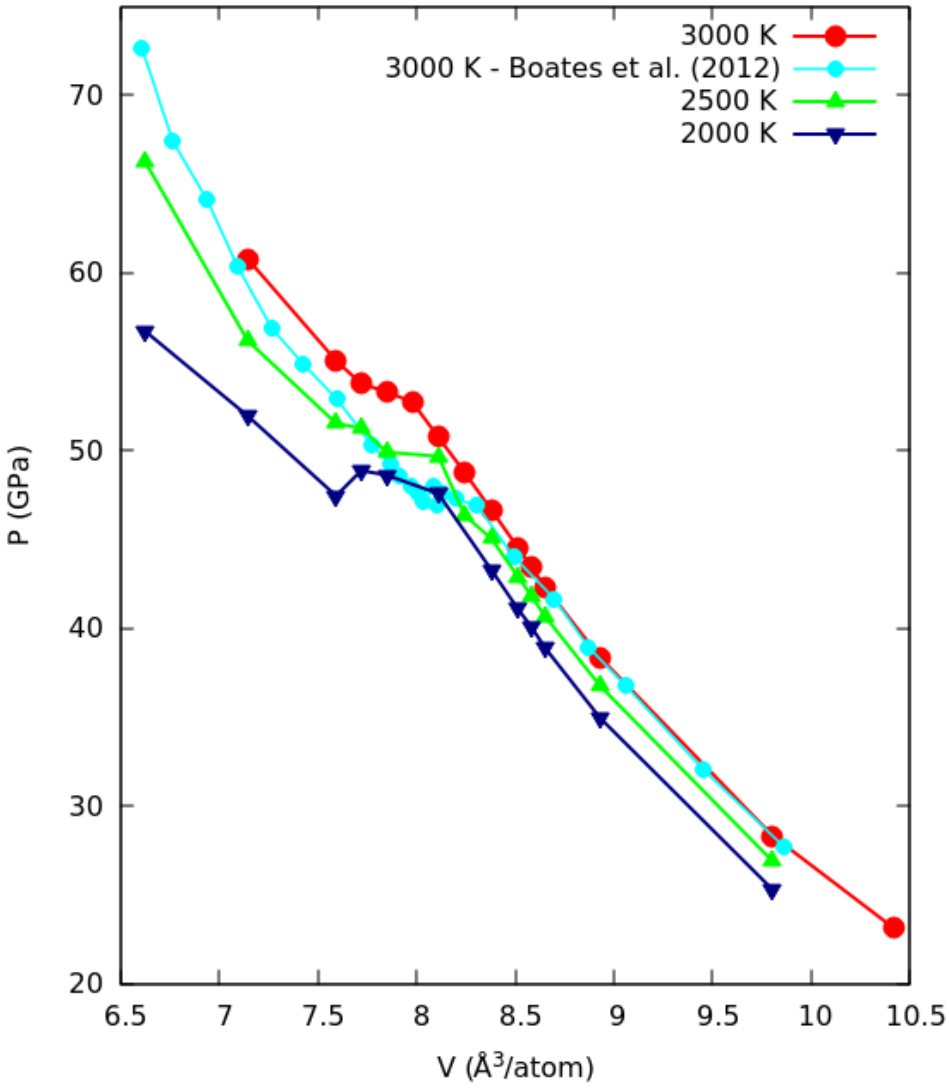


Figure 5.1: Pressure-Volume relationship at three different temperatures (2000K, 2500K, 3000K), along with the results obtained by Boates et al. (2012) [7]

better show the situation.

As can be expected, we find that the first two nearest oxygen to carbon atoms are within a given radius of the carbon atom. The situation becomes more complicated for the third and fourth nearest oxygen to a carbon atoms: we observe in both cases two main peaks (for each distribution), the first one - below 1.75Å - corresponds to bonded oxygens, while the second peak corresponds to non-bonded atoms. This indicates that the polymerization will form CO_3 and CO_4 atoms. The potential issue is that there is a significant overlap between the two distributions which implies that the use of a cut-off will create to type of errors: some atoms that are bonded will be counted as not-bonded when thermal fluctuations pull them apart at a distance longer than the cut-off; atoms that are not bonded but merely colliding will be counted as bonded if they get within a

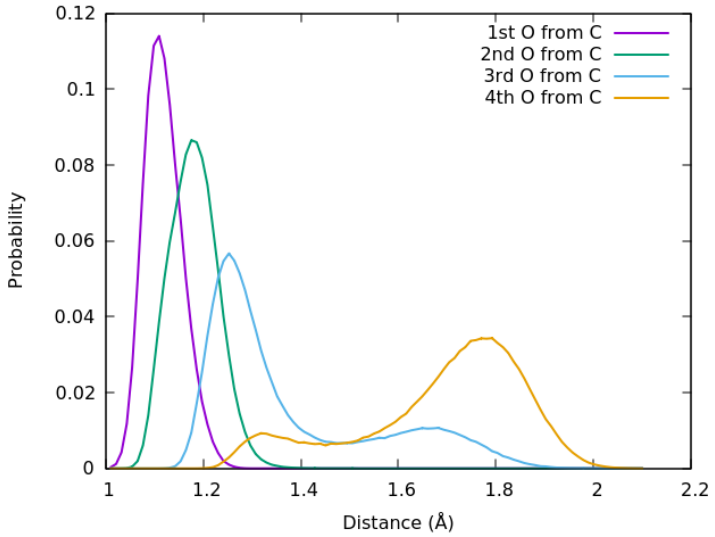


Figure 5.2: Distribution of the distance of the four closest oxygens to carbon atomsin the polymeric conditions (65 GPa and 3000 K)

distance smaller than the cut-off value.

In Figure 5.3, wee see a similar situation for the distribution of distances of first carbon atoms to carbon atoms. In this case the bonded peak is much smaller, indicating that C-C bonds will be rare. As the height of the bonded peak is comparable to that of the overlap between it and the non-bonded one, the choice of the cut-off may largely impact the number of C-C bonds that will be counted. We note however, that the main peak mosty remains still close to the center carbon, indicating that although the carbon atoms may not be directly connected, it is likely that they have at least one oxygen atom in common.

Finally, the partial pair correlation function related to the O-O distances clearly indicates that there is no O-O bond, and therefore we can ignore the cases of O-O bonding for the rest of the analysis of the liquid phases.

The next step to asses the validity of cut-offs would be to count the number of errors that one would make for a given cut-off using a reference method that would be reliable enough to be used as truth value. However, there is no perfect indicator to determine whether or not two atoms are bonded.

We therefore decided to use two different methods each relying on different principle. First we used a criterion based on the Electronic Localization function (see 1.1.3) which makes use of the very accurate precision of the *ab initio* calculation. We then used a method based on Density Peak Clustering (cf 3.2.3) to determine the coordination number of carbon atoms, and therefore the existence of bonds, making use of our large data sets.

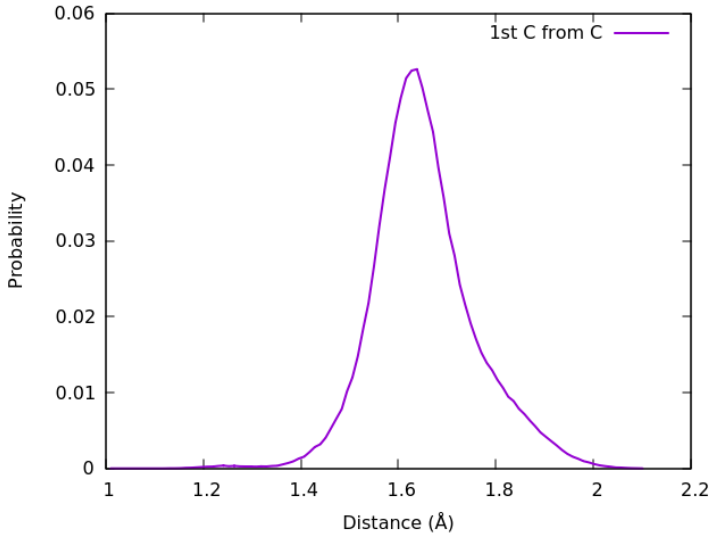


Figure 5.3: Distribution of the distance of the nearest carbon to carbon atoms in the polymeric conditions (65 GPa and 3000 K). The bonded peak is barely visible between 1.2 and 1.3 Å.

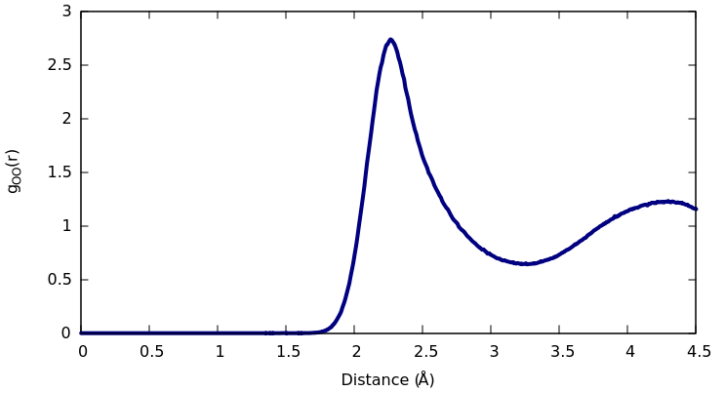


Figure 5.4: Partial pair correlation function for the O-O distances.

5.4.1 ELF in the middle

The first approach makes uses of the Electron Localization Function (see1.1.3). The appeal here is that it provide an electronic criterion to determine bonding. In order to construct our criterion we first looked at the values of the ELF along the path from each carbon atom to each of its bonded counterpart (using the distribution of Figure 5.2 to consider atoms separated by distances within the bonded peak). The results are shown in Figure 5.5 for atoms of carbon and oxygen closer than 1.75Å from each other.

In order to proceed we computed the ELF for 1000 frames of simulation at the most extreme conditions (65 GPa, 3000 K) using CPMD[157]. We obtained a 60x60x60 grid for a cubic cell of side 8.82Å, providing a precisin of 0.174Å in each direction. As mentionned previously, there is no O-O bond in our simulations, but we also did not find any C-C bond in the 1000 frames that were selected for analysis with ELF therefore we will focus

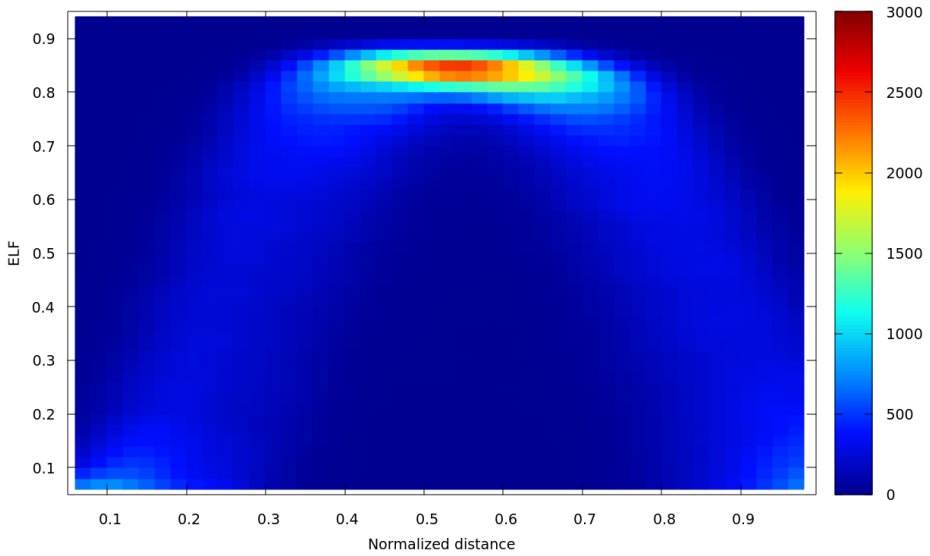


Figure 5.5: Heatmap of the value electron localization function along the (normalized) distance between carbon (located in 0) and oxygen (located in 1) atoms that are closer than 1.75Å from each other

on C-O bonds in our analysis.

In figure 5.5 a clear maximum value of the ELF value exists almost in the middle of the distance between the carbon and oxygen atoms. This indicates an increased localization of electrons at the same place, which fits nicely with the concept of a covalent bonding. Interestingly, the maximum is slightly displaced toward the oxygen, which is likely due to the fact that oxygen is more electronegative than the carbon and in average pulling the covalent bond closer to it.

Although we have indentified a common behavior of the ELF along the distance for bonded atoms, we wanted to check the behavior of this metric in the cases of non-bonded ones. To do so, we carried out the same analysis with atoms further than 2.0Å to each other, insuring that they would not be bonded (figure 5.6) .

As expected, a different behavior emerges here: the distribution shows two peaks of the ELF along the distance, one close to each atom, and a local minimum appears where a maximum was observed in the bonded case. Here, it does seem that the electrons are localized around their respective atoms. Interestingly, we see that the first peak, around the carbon atom is much lower than the second one, related to the oxygen atom.

Although we did not find any C-C bond, we decided to check whether the same general behavior is observed for non-bonded carbon-carbon first neighbors as in the C-O case (figure 5.7).

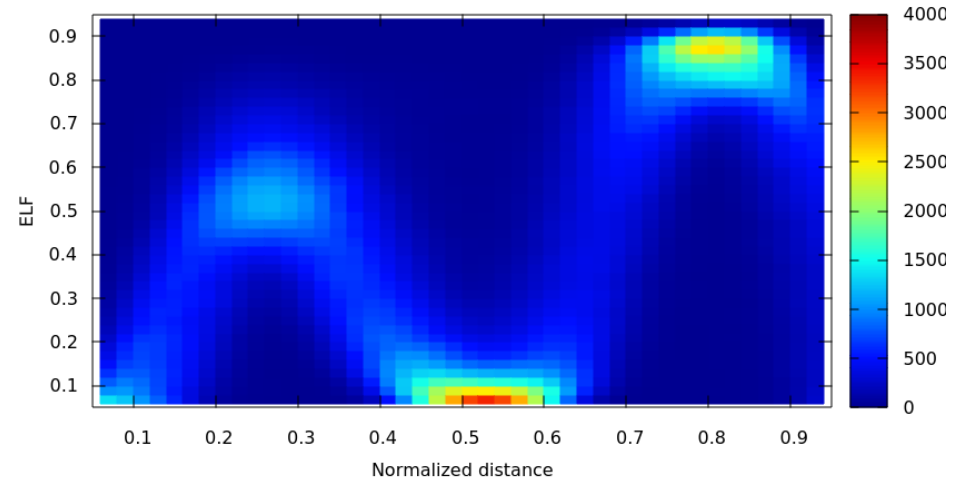


Figure 5.6: Heatmap of the ELF along the (normalized) distance between carbon and oxygen atoms more than 2.0Å away from each other. Carbon is in 0 and oxygen in 1.

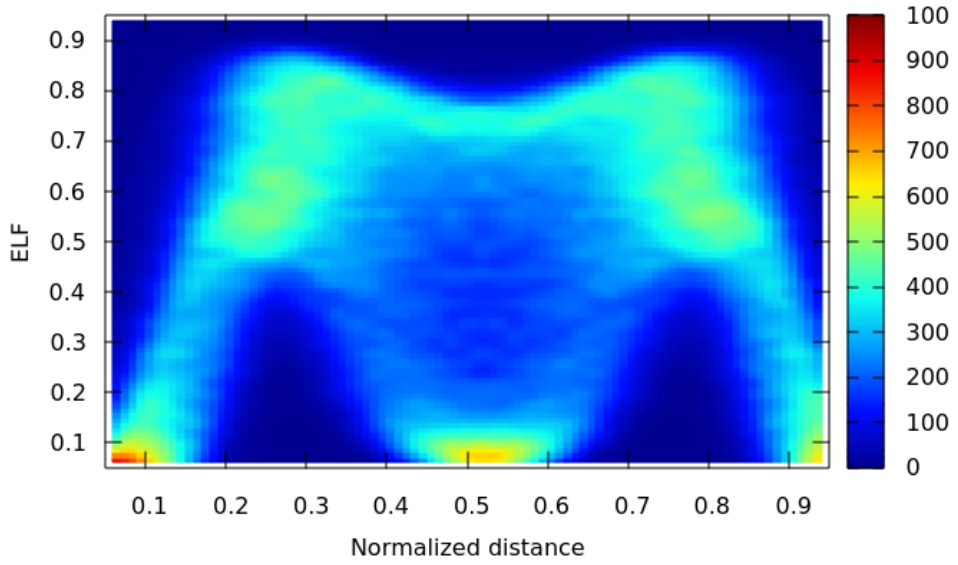


Figure 5.7: Electron Localization along the distance between two carbon atoms more than 2.0Å away from each other

The results here (figure 5.8 are more complex than in the C-O bond: although we do have a similar trend of two clearly established peaks close to each atoms, in the intermediary region, two distinct behavior can be observed: we can either have the same low ELF value minimum (ELF∼0.15) or a high ELF value minimum (ELF∼0.6) which seem to match neither a bonding state nor a complete non-bonding one.

As carbon atoms are relatively far from each other in the simulation associated with those results (more than 2Å away) we thought that this effect might be due to the presence of a shared oxygen atom between the two carbons. In order to check this, we repeated the analysis, selecting only the cases where two carbon shared a common oxygen atom (figure 5.8).

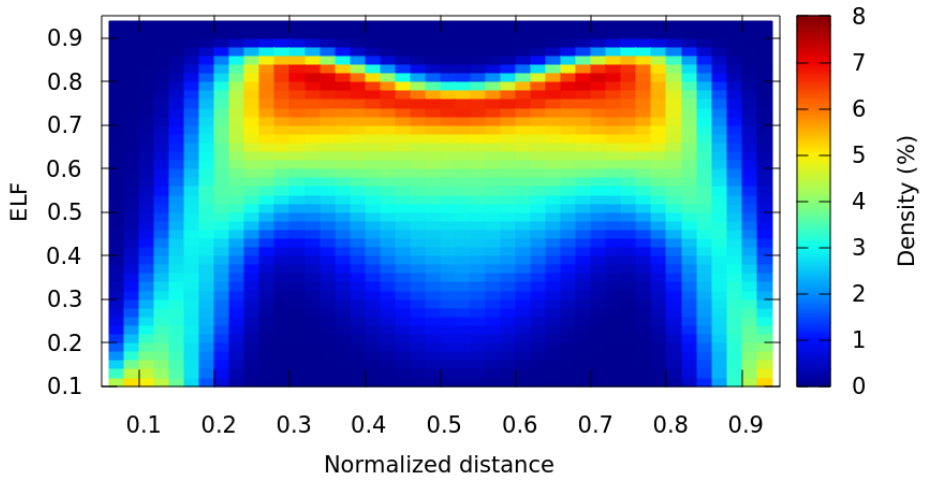


Figure 5.8: Electron Localization along the distance between carbon atomsthat share an oxygen as neighbor. Distances are takent between 2.0 and 2.8Å

The results (figure 5.8) confirm the hypothesis as we clearly see the expected behavior: two peaks, one for each atoms and a high value ELF (~ 0.6) minimum close to the middle of the C-C (normalized) distance. The conclusion here is that it is likely that the ELF related to the orbital of the shared oxygen atom somewhat extend up to the middle of the C-C distance. The conclusion here is that although the ELF behavior along the distance from one atom to another may yield clues about their bonding state, this information will also be affected by the local environement, which therefore should also be taken into consideration.

The conclusion we drew from thos results is that it seems that it is possible to determine whether or not carbon and oxygen atoms are bonded using the ELF value in the middle of the C-O distance. From the previous results, it seemed that using a cut-off value of 0.75 on the value of the ELF at the middle point between atoms may be used effectively to determine whether a covalent bond existed between the two. In the rest of this work, we will refer to this method as the *ELF in the middle* method, as a short hand.

In order to test the validity of the use of cut-off, we computed the distribution of the ELF in the middle point of C-O distances for distances between 1.0 and 2.5Å (figure 5.9). The results shows encouraging results as we see that most of values of ELF superior to 0.75 tends for distances between 1 and 1.75 Å. In general, the graph reflects the same kind of information as the distance distribution 5.2: two regions are well defined

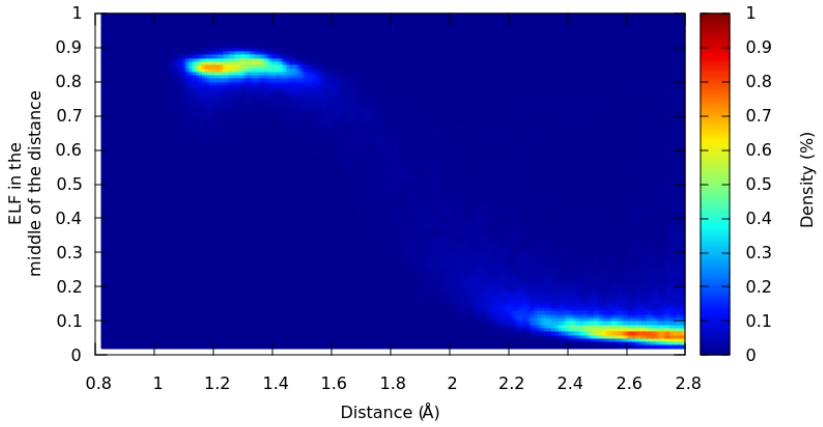


Figure 5.9: Electron Localization Function in the middle of C-O interatomic distances as a function of the distance1 (for distances below 2.5Å). **Top:** distances between 0.8 and 2.5Å. **Bottom:** between 1.5 and 2.1 Å

corresponding to bonded (distances$d < 1.6$ Å and ELF > 0.75) and non-bonded states (distances$d > 2.0$ Å and ELF < 0.4), while in between a intermediary, more difficult region to define exists.

Using the ELF cut-off of 0.75 as truth value to determine bonding, we find that using a cut-off of 1.75Å results in 6.22% of errors, with 0.03% coming from missing bonds that exists and 6.18% by predicting bonds that do not exists. We consider that this value is low enough that the cut-off may be used at least for general structural considerations, but it may require more advanced methods to understand aspects based on the dynamics or kinetic aspects of the system such as lifetimes of molecules.

5.4.2 Unsupervised learning

Although the ELF in the middle method provides a good criterion for bonding, as we saw in the case of C-C interactions, it can be affected by the environment of the atoms. In this section we show a method that aims at determining bonding between atoms by specifically accounting for their environment.

The idea is the following: one can describe roughly the local environment of an atom by using the sorted distances of its N first neighbors of a given type as described in (2.2). As we know that we do not have C-C or O-O bond, we decided to focus on the first 5 oxygen neighbor of carbon atoms in order to do so.

Using this description of each atoms, we used Density Peak Clustering (3.2.3) to group together carbons with similar environment. In order to compare two carbon atoms, we took the Euclidian distance of vector so created. That is, for two carbon i and j, their distance $v_{i,j}$ was defined as:

$$v_{i,j} = \sqrt{\sum_{k=1}^{4}(d_{i,k} - d_{j,k})^2} \tag{5.1}$$

where $d_{i,k}$ is the distance of carbon i to its k^{th} first oxygen neighbor. We tested the DPC algorithm for various values of the d_c parameters (cf. 3.2.3) to check the validity of the classification and we used the same data set than the one used in the *ELF in the middle* case, in order to be able to compare the two results.

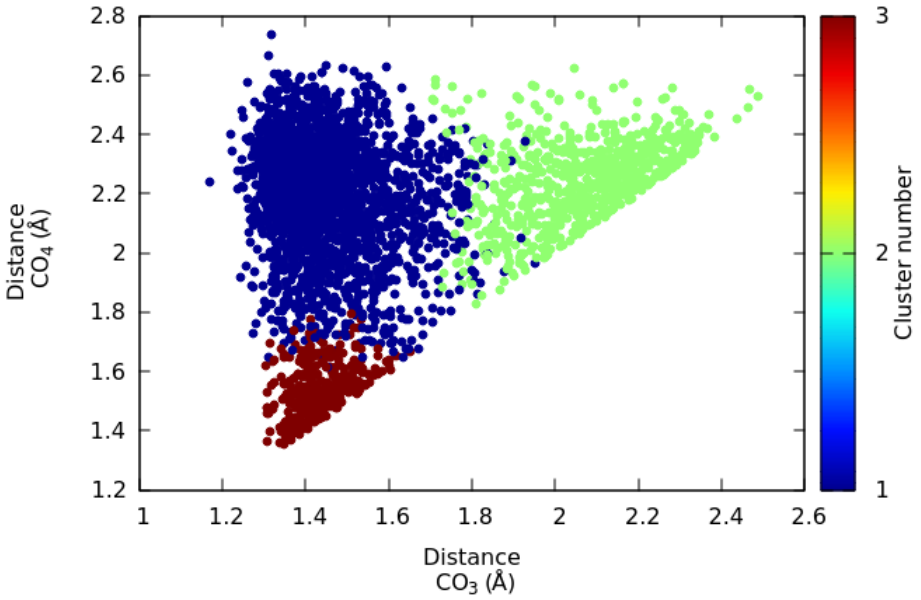


Figure 5.10: Distances to the third and fourth oxygen atom for all carbon atoms in a 1000 step trajectory colored by their label affected by the DPC algorithm.

The result of this analysis is shown in figure 5.10. We see that DPC finds three different clusters, one for each of the coordination number observed for carbon in those experimental conditions (2,3 and 4).

We observe that cluster 2 gathers carbons that have both most of their 3^{rd} and 4^{th} carbon atoms at distances above 1.75Å and therefore will be associated with carbons with a coordination number of 2. Cluster 1 on the other hand, regroups carbons that are likely bonded to their 3^{rd} nearest oxygen atom but not their 4^{th} nearest. Finally, cluster 3 contains carbon that are likely bonded to their four first nearests oxygen atoms.

Once trained, the clustering algorithm may be used to predict the coordination number of a carbon atom by assigning it to a given cluster based on its distances with the carbons from the training set. From the coordination number it is easy to go back to the bonds that a given carbon atom has with its neighbor, and as the system contains only

C-O bonds, to the bonds of all the system (assuming that if an atom has a coordination number of 3, it is bonded to its three closest neighbors). Although the procedure should be repeated on a different set featuring C-C bonds if one wants to be able to predict the correct coordination number in simulations where those bonds appear.

The interest of using this metric is that it relies on the statistical distributions of the four first distances of oxygen for each carbon in the data set and therefore takes into account the local environment to regroup similar atoms and makes use of the amount of data provided in order to do so.

Using the results of this method as truth value, we can also compute the errors that would result from using a cut-off to determine bonding. In this case, for a cut-off at 1.75Å, we get an error of 3.75%, 1.93% coming from counting bonds that do not exist while 1.18% come from not counting bonds that do. Overall, we see that the cut-off here is a very good match and works relatively well. This results also confirms that taking into account local environment only marginally improves on the simple use of single distances to determine bonding.

5.4.3 Combining ELF and unsupervised learning

Finally, we show that we can combine both methods can be combined together to take into account both the local geometry and an electronic criterion. In order to do so, we can first simply plot the ELF value at the middle point as a function of the distances between carbon atoms and their nearest oxygen neighbor .

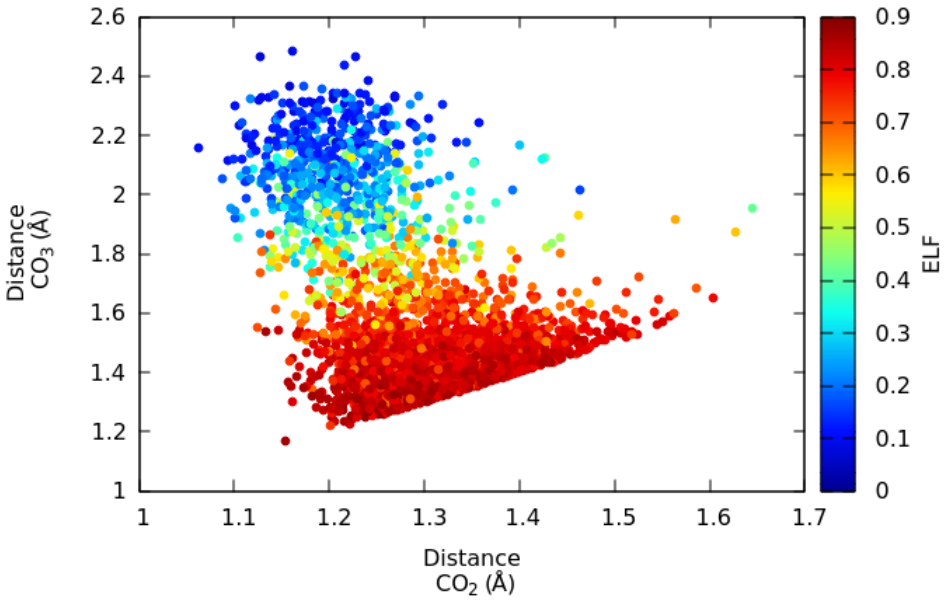


Figure 5.11: Electron Localization along the distance between carbon and oxygen atoms

In the figure 5.11, we focus on the ELF value of carbon atoms with their third clos-

est oxygen neighbor, which we compare the its distances to the second and third carbon atom. We see that globally the value of the ELF at the middle point between atoms mostly varies with the distance to the third oxygen neighbor, however the influence of the distance to the second nearest carbon atom is visible. Therefore, in this case, it does seem that a cut-off that depends not only on the distance between the carbon and the third closest oxygen atom but also slightly on the distance with its second nearest oxygen will be necessary.

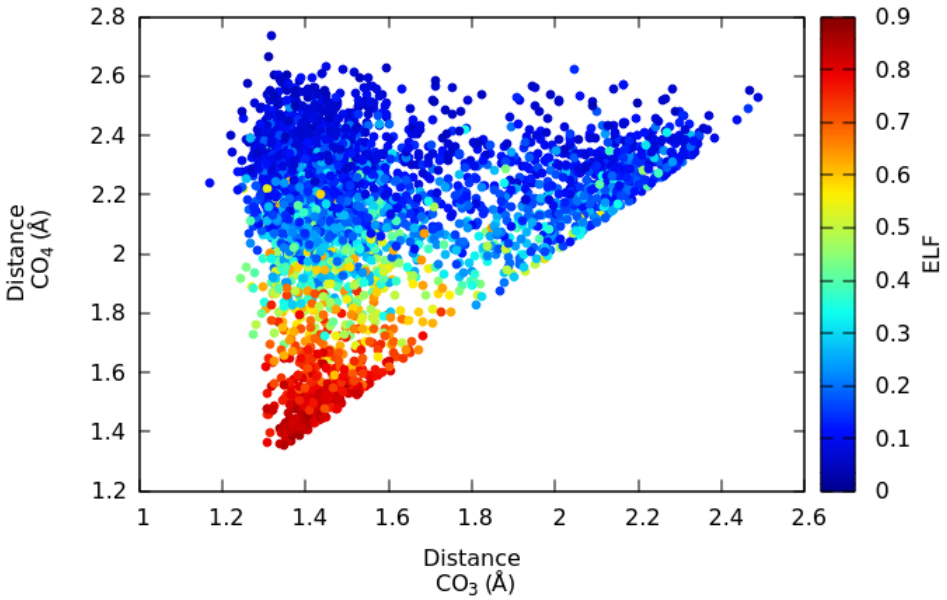


Figure 5.12: Electron Localization Function in the middle of the bond between carbon atoms and their fourth nearest oxygen as a function of the distances between the carbon atoms and their third and fourth nearest oxygen

As for the ELF values in the middle of a carbon atoms and their fourth nearest neighbor (figure 5.12) we see that it does seem to depend only on the distance between those atoms. This seem to indicate that the use of a cut-off using only the distance between carbon and fourth nearest oxygen is likely to be efficient if the cut-off if chosen properly. Overall both figure 5.11 and figure 5.12 validate the use of single cut-off to determine bonds between carbon and oxygen atoms.

Finally we used the Density Peak Clustering algorithm directly on the values of the ELF at the middle point between carbon atoms and their first four nearest oxygen 5.13. In this case as well it is possible to identify the three coordination number for carbon atoms (2,3,4 neighbors), and DPC is able to indentify them for any reasonable choice of d_c (see 3.2.3).

If now we were to plot the same point with the same affectation but in the distance space (figure 5.14) we see results that are closer to those obtained by using DPC on the distances rather than those obtained solely with the value of the ELF in the middle of

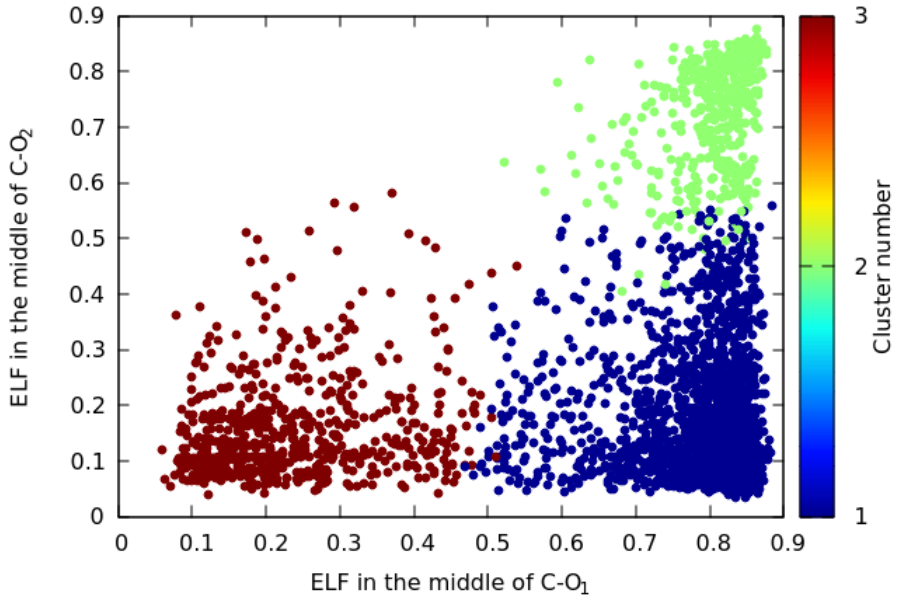


Figure 5.13: Electron Localization along the distance between carbon and oxygen atoms

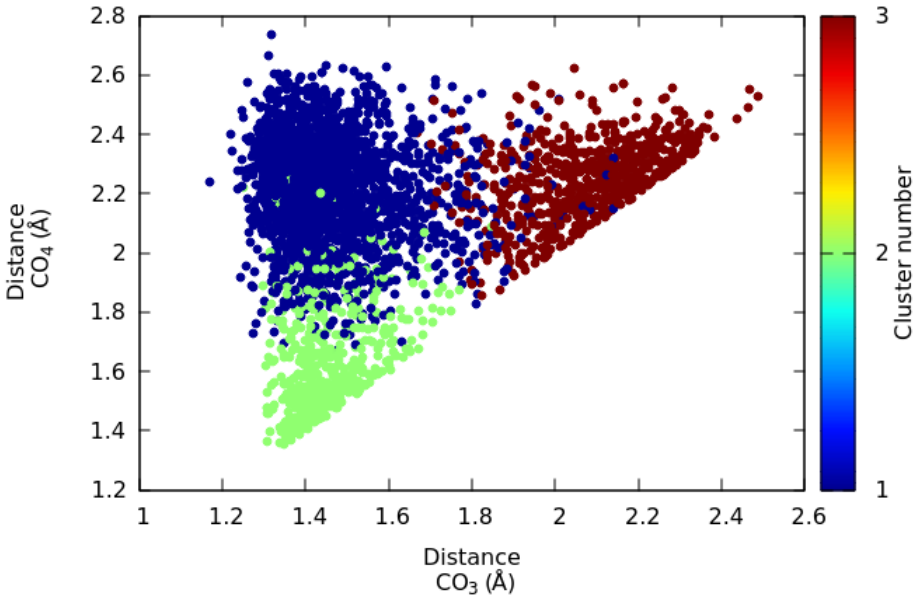


Figure 5.14: Electron Localization along the distance between carbon and oxygen atoms

distances. In general the disagreements mainly seem to occur within the intermediate region and are relatively minor.

In this section we have therefore shown that regardless of the method one uses to determine bonding, the use of cut-offs is legitimate, at least for structural analysis, and results in relatively low ($\sim 5\%$) percentage of errors. From this point on, we will therefore use a cut-off $d_{cut} = 1.75$Å to proceed for the structural analysis.

It is likely however, that the dynamics of the atoms in the system, that is, their evolution from one chemical state to another may not be portrayed accurately enough by cut-offs (due to the systematic errors and the flickering phenomenom, where atoms may move back and forth around the cut-off value) and we will likely need to use more involved methods to do so.

Although it is likely that the systematic use of either of the two methods (clustering or ELF in the middle) would have resulted in a more accurate portrayal of the chemical structures in the system, however both methods turned out to require long analysis time[1] and we therefore thought that they would be excessive with regard to the correction that they brought.

We also carried out similar *ELF in the middle* analysis on a prebiotic chemistry system, provided by a post-doctoral research of the team, Andrea Perez-Villa, with similarly interesting results (see **??**).

5.5 Coordination numbers

Using the cut-off determined above, we can easily compute the coordination fraction of carbon and oxygen atoms, in order to gain some information about the transformation, in the same spirit as [7]. As we have noted above, we expect only to have twofold, threefold and fourfold coordinated carbons and we have therefore we focused solely on those (figure 5.15). We did note however, a suttle rise of carbon with only one oxygen neighbors at the highest temperature and pressure, but the fraction remains extremely small (less than 0.1% at 3000 K and 65 GPa).

The main element that is readily available is the decrease of twofold coordinated carbons (C_2) with increasing pressure (figure 5.15). However, contrary to what was previously reported [7], the decrease seem to occur in two steps: first a slow decrease that is highly temperature dependant and start at pressures as low as 25 GPa at 3000 K, followed by a faster one around 48 GPa at 2000 K and 56GPa at 3000K (figure 5.15). The general trend is in agreement with the results of Boates et al [7], except for the shift in pressure discussed in 5.3.

The first and very progressive decline of C_2 (occuring mainly in the 20-48GPa range at 2000K, up to 55GPa at 3000K) is attributed to the formation of short-lived C_2O_4 or C_3O_6 molecules, which is favored both by increasing temperature and pressure. Although this was also observed in [7], it is slightly more pronounced in our case. We note that this results matches the observation of dimerization of carbon dioxide at high temperature [58], or in an amorphous phase around 150K [59], where similar small chains where

[1] In the case of ELF, it also would have required to redo most of the calculations to compute the ELF, which would have been extremely costly, both in termes of computation time but also problematic in terms of storage.

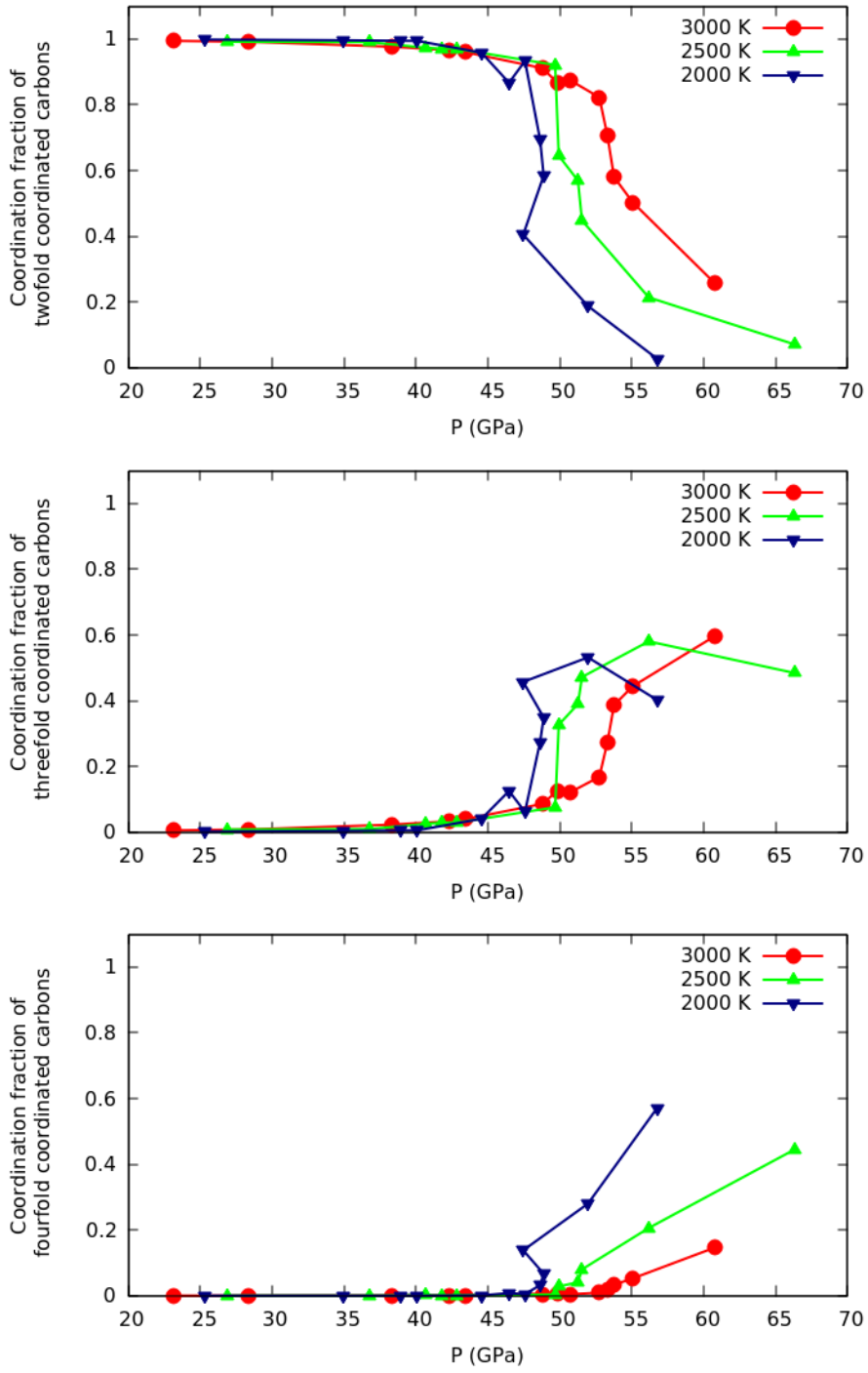


Figure 5.15: Coordination fraction of twofold (top), threefold (center) and fourfold (bottom) coordinated carbon as a function of the temperature and pressure.

observed. The fact that those molecules are short lived yet still significantly affect the average number of C_2 in the simulation indicates that they from relatively frequently. This finding may indicate that the molecular fluid close to the transition (and at high temperature in general) is more reactive than previously reported.

The second and sharpest drop was previously interpreted as the signal of the transition between the molecular and polymeric regime [7]. Our data concur with this explanation as this diminution corresponds to a strong increase first of threefold (C_3) coordinated carbons, folowed in a second step by fourfold (C_4) coordinated ones. Both steps are highly temperature dependant: at 2000 K and 60 GPa the first step is already passed and C_4 are more common than C_3 while at 3000 K and 60 GPa the first step is still underway and C_2 are still more frequent that C_3 and C_4 combined.

Interestingly, it is likely that the behavior at 2000 K and 3000 K may be largely different as the former seems to indicate a very fast transition to a polymeric liquid dominated with C_4 carbons while at 3000 K the polymeric liquid is dominated with C_3 carbons.

This consideration is further confirmed when considering the fraction of oxygen with two covalent bonds (O_2) as in figure **??**. Where we see an important rise in O_2 at the molecular-polymeric transition pressure, especially important at 2000K. We note that at 56GPa and 2000K, 80% of the oxygens have two neighbors, indicating that the underlying network is almost completely connected. At 3000K however, the fraction of O_2 barely reaches 50% at 62 GPa, which indicates a looser network.

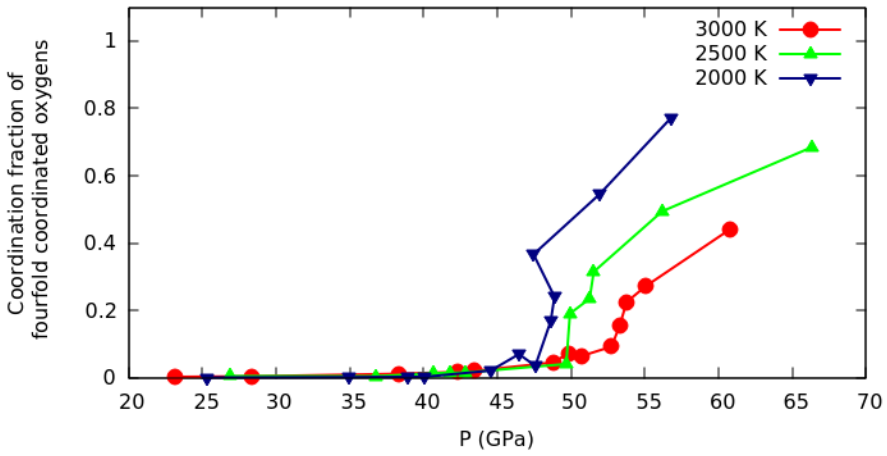


Figure 5.16: Coordination fraction of oxygen with two neighbors over the pressure-temperature range for the liquid-liquid transition

All things considered, by considering the coordination number of oxygen and carbon we identify clearly three fluid behavior: the standard molecular fluid at low temperature and pressure, a molecular liquid where there is a formation of small reactive chains and a polymeric liquid that appears from a rapid diminution of CO_2 units in the system, potentially with different behavior at high and low temperature. We expect the reactive molecular liquid to occur mainly between 20 and 45 GPa - although we expect that at 2000 K, the behavior actually appear over 35 GPa - and the polymeric liquid appears over 48GPa, although this transition pressure is highly temperature dependant.

5.6 Reactive molecular fluid

In order to study the molecular liquid region where there is a formation of short chains, we first verified that the formation of the chains were not an artefact due to the cut-off. We used the ELF to study the formation, life and decay of an identified C_2O_4 dimer and C_3O_6 trimer and looked for evidence of bond creation/destrution. Doing so, we were able to identify a dimerization mechanism shown in figure 5.17 for C_2O_4 dimer and confirm the existence of bonds in the C_3O_6 trimer (figure 5.18).

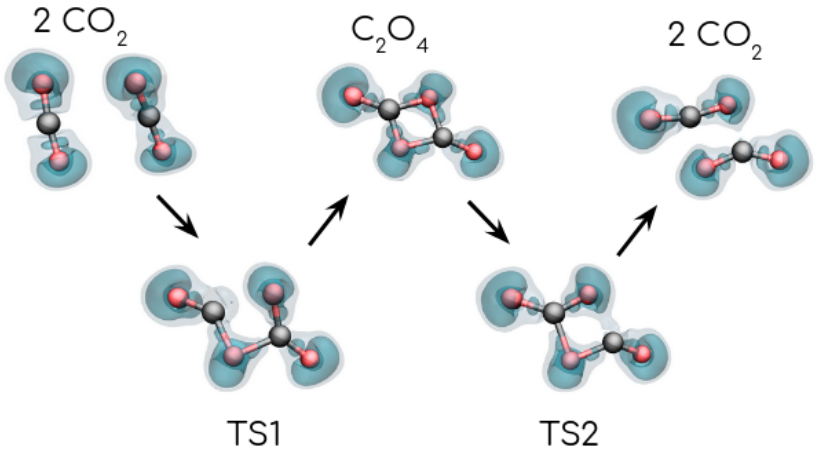


Figure 5.17: Dimerization mechanism with ELF isovalues (0.6 in ble and 0.8 in grey)

Almost all short chains that formed during the simulations were intermediate states of the dimerization process (TS1 and TS2 in figure 5.17). The dimer proper was found to be up to four times less frequent and the trimer was exceptionnaly rare, except in the 40-48GPa region. Interestingly, the trimer does not seem to form from an incomplete dimer but by simultaneous assembling of three molecules within 10fs. We note that the dimers could allow for the exchange of oxygen between the original CO_2 molecules.

In order to compute the limit of this reactive molecular liquid, we computed the fraction of carbons that are part of C_2O_4 chains over the pressure-temperature range of interest (20-45GPa, see figure 5.19). Interestingly it seemed, that even at pressures as low as 25GPa, the fraction was on the order of 1% both at 2500 and 3000 K below 25GPa. We note that this number was not reached at 2000K before 40 GPa. Overall, the phenomenom increases with both pressure and temperature.

This is particularly interesting as it was previously suggested that the C_2O_4 dimers may actually hinder polymerization [58] and even be a precursor to dissociation [7]. However, the fact that pressure also increases suggests otherwise. Furthermore, we note that many dimer-like units where observed in the complex chains of the polymeric fluid at 3000K, casting more doubt about this hypothesis.

It is likely however, that those small structures are a sign of the weakening of the C-O double bond and a change in the behaviors of CO_2 molecules, from purely long

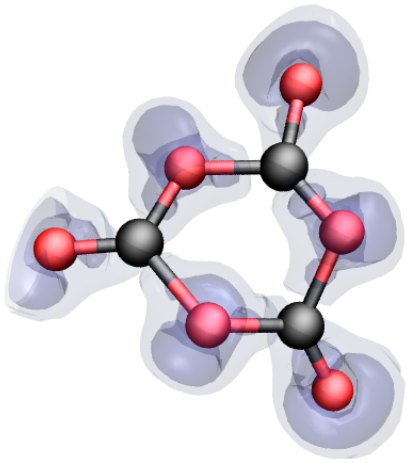


Figure 5.18: C_3O_6 cyclic trimer with ELF isovalues (0.8 in blue, 0.6 in grey)

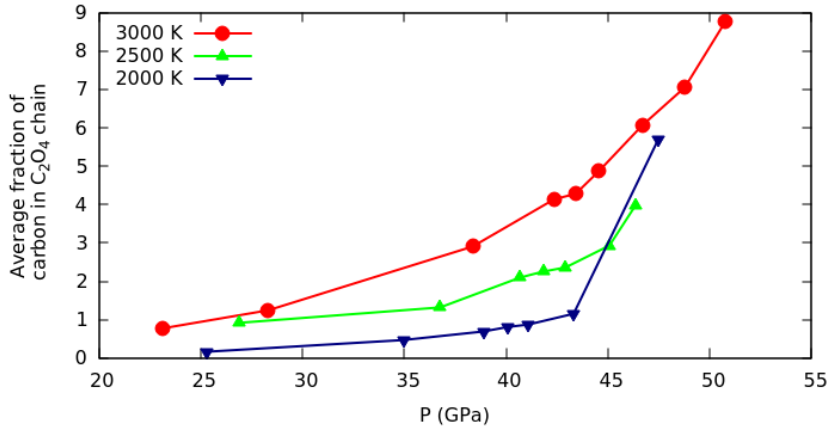


Figure 5.19: Fraction of carbon atoms belonging to C_2O_4 chains during a simulation as a function of the pressure and temperature

range interactions to a reactive behavior. Those reactions, depending on the experimental conditions, may push the system into either dissociation (at high temperature) or polymerization (high pressure).

Experimentally, those results may be confirmed by experiments through raman spectroscopy: indeed, according to [58], the C_2O_4 dimer has an raman activity around 2300cm^{-1}, although we note that as none of the observed events had a lifetime over 1ps, it is likely that the signal may be difficult to detect experimentally.

The potential existence of a reactive molecular phase is of particular interest for geological application as it seems that its domain interesects the geotherm in the 40-45 GPa region at 2000 K. Those experimental conditions corresponding to the depth where recents experimental studies suggest that CO_2 may form [4, 5] from interactions between carbonates and SiO_2. This would imply that the formed CO_2 would be under a reactive

liquid form that would react very quickly with its surrounding environment and therefore play a role in the chemistry of the lower mantle.

5.7 Polymeric liquids

As shown in (5.5), the polymeric liquid forms from the molecular fluid above 48 GPa and is characterized by the formation of CO_3 or CO_4 units, which form complex polymeric chains. To characterize the progressive formation of those chains with increasing pressure, we computed the distribution of the sizes of molecules at 2000 K, as in figure 5.20.

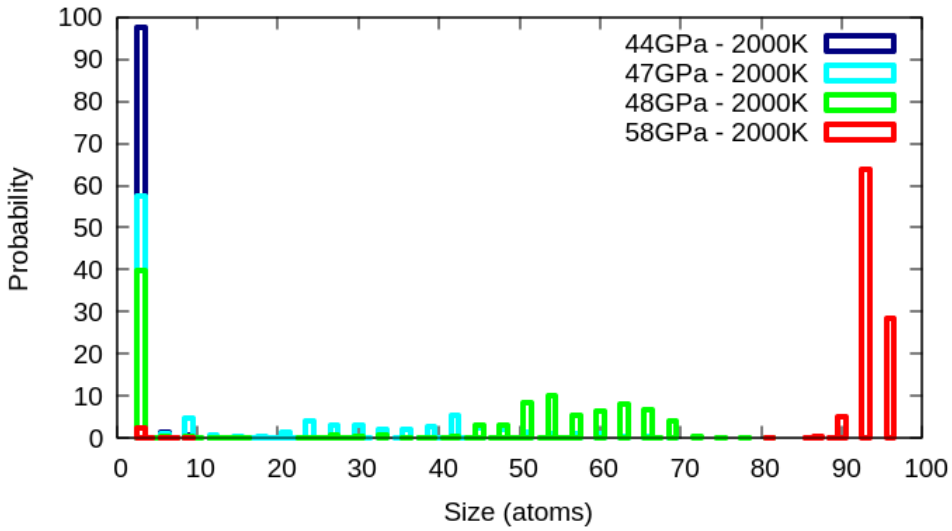


Figure 5.20: Distribution of the size of molecules at 2000 K at 44, 47 and 48 GPa.

In figure 5.20, which shows the evolution of the distribution of molecule size (in atoms) as a function of the pressure at 2000 K, we see that the transition between a molecular liquid, with marginal formation of small chains such as C_2O_4 and C_3O_6, to a polymeric liquid, which is characterized by long chains, happens very quickly between 44 GPa and 47 GPa. Indeed, although at 44 GPa, there is only a very small amount of C_2O_4 molecules and an overwhelming fraction of CO_2 molecule (figure 5.20), we see that at 47GPa, the system has already transited to a system where long chains containing between 7 and 14 CO_2 units are stable, while the CO_2 fraction dropped below 60% (figure 5.20). At 48 GPa, this fraction drops to 40% and the large molecules are now between 21 and 78 atoms long (figure **??**). At 58 GPa the system is already fully polymerized, as is evident by the very low fraction of CO_2 molecules (below 5%), and the molecules that are most present are composed of 27 to 32 CO_2 units (figure 5.20).

This evolution (figure 5.20) indicates a very fast polymerization and the formation of large molecules that, according to the results of figure 5.15, are mostly composed of CO_4 tetrahedra, which is confirmed by visual observation of the trajectory, where the

formation of a relatively stable three-dimensional network is observed.

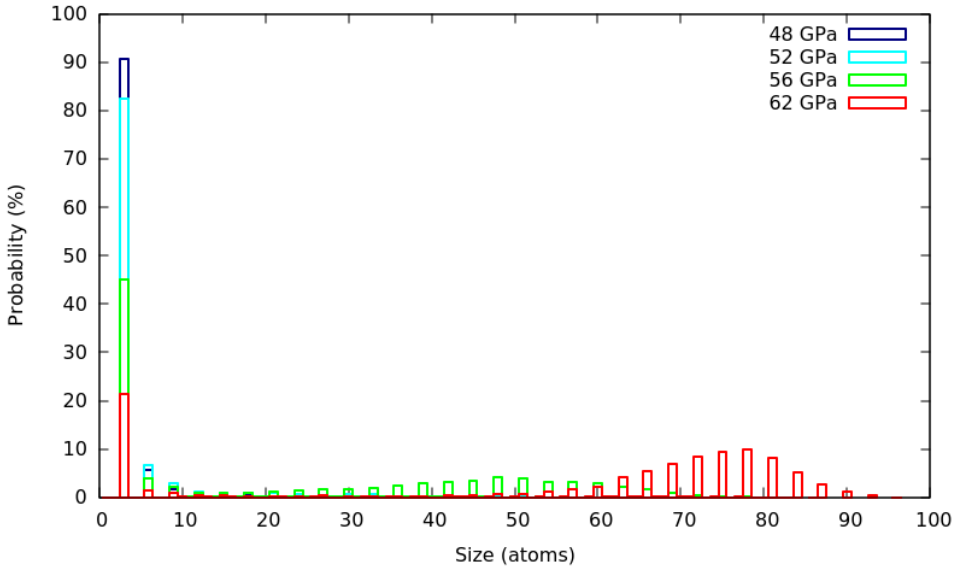


Figure 5.21: Distribution of the size of molecules at 3000 K at 44, 47 and 48 GPa.

On this other hand, we have a slower polymerization at 3000K (figure 5.20). The average size of the molecules increases progressively between 40 and 65GPa (figure 5.20). At that pressure, although we observe the formation of large chains, we have not yet reached a fully polymeric liquid (figure 5.20). Furthermore, the spread of the size distribution, in this case, is still relatively wide and the fraction of CO_2 units decreases more slowly than at 2000K (figure 5.20).

We explain this by the nature of the polymeric liquid, which at this experimental conditions, is observed to be highly dynamic and where CO_2 molecules are being exchanged between the various large molecules (or between different part of the same molecule) to stabilize their unstable ends which often feature likely unstable COO units.

The distribution of molecular sizes (figure 5.20) is coherent with both the observed behavior of the liquid and with the fraction of carbon coordination fractions. The high and low-temperature liquids indeed seem to exhibit different behavior between a stable three-dimensional network that forms quickly from the molecular fluid to a complex system of large molecular chains at high temperatures.

Those observations are confirmed by the partial pair correlation functions at 2000, 25000 and 3000 K (between carbon and oxygen atoms) as shown in figure 5.22. In this figure, we see that the first peak is composed of two contributions: one for the CO_3 units and one for the CO_4 ones. The CO_3 contribution is the one corresponding to distances around 1.2 angströms while the CO_4 contribution corresponds to the peak around 1.4 angströms. Indeed, we expect at least some CO_3 units to still contain double C-O bonds, while CO_4 units should exclusively contain longer single bonds. We observe that the

contribution of CO_3 units is much stronger than the CO_4 one at 3000K, the latter being almost negligible in comparison. At 2000 K however, the situation is reversed: the CO_3 contribution is visible but weak compared to the CO_4 one. Furthermore, the secondary peaks of the pair correlation function are much more defined at 2000 K than at 3000 K. This indicates that the overall structure of the network formed at 2000 K is much more stable than at 3000 K, which coincides with direct observation of the simulation.

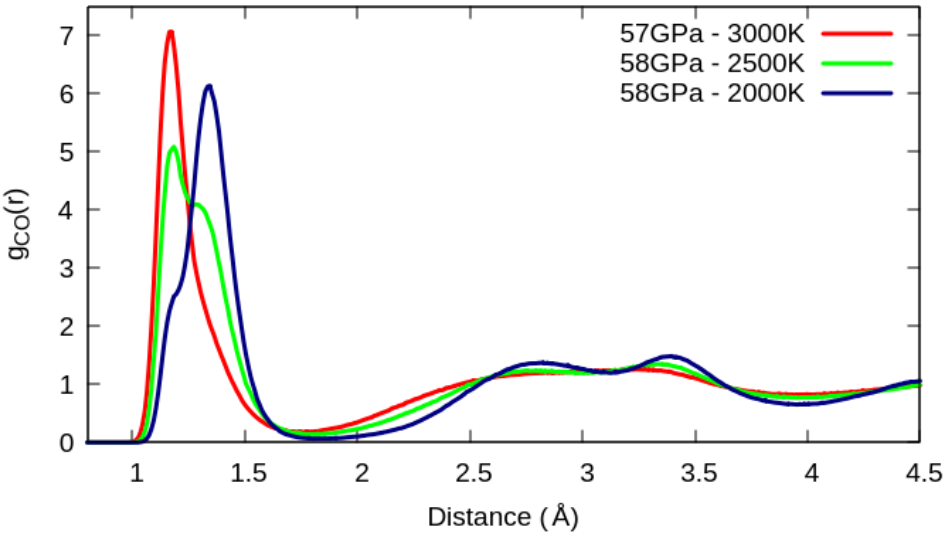


Figure 5.22: Pair correlation function of carbon and oxygen

To confirm the observation of low diffusion in the low-temperature polymeric liquid, we computed the diffusion coefficient over the whole pressure-temperature range (figure 5.22) using the mean square distance (MSD). We observe a general diminution of the diffusion coefficient with increasing pressure and an increase with temperature. We also see a drop of the value of the diffusion coefficient at 2000 K, to the point where the diffusion coefficient is almost 0 above 50 GPa. This marked drop is likely an indication of an amorphization of the polymeric liquid into an amorphous polymeric liquid. Interestingly, all results about the 2000 K liquid bear a strong similarity with the low-temperature amorphous phase of carbon dioxide observed in [59] via compression of CO_2-III.

We also computed the difference between the MSD of the oxygen and carbon atoms (figure **??**, in order to check that the mechanism of diffusion of both type of atoms where the same. The results (figure **??**) shows that in the molecular liquid, the oxygen have a somewhat higher diffusion than carbon atoms, likely due to the possibility to rotate around the central carbon atom, however this difference quickly becomes constant. On the other hand, we see that at 2000 K and 58 GPa, corresponding to the amorphous phase, there is no difference in the diffusion of carbon and oxygen, as can be expected as we observe almost no movement of atoms in this system. At 3000 K and 58 GPa, however, we see that oxygen atoms diffuse faster than oxygen, and that this difference in MSD steadily increases with time. This is likely due to a relatively large number of exchange of oxygen during the association/dissociation of CO_2 units from the larger molecules.

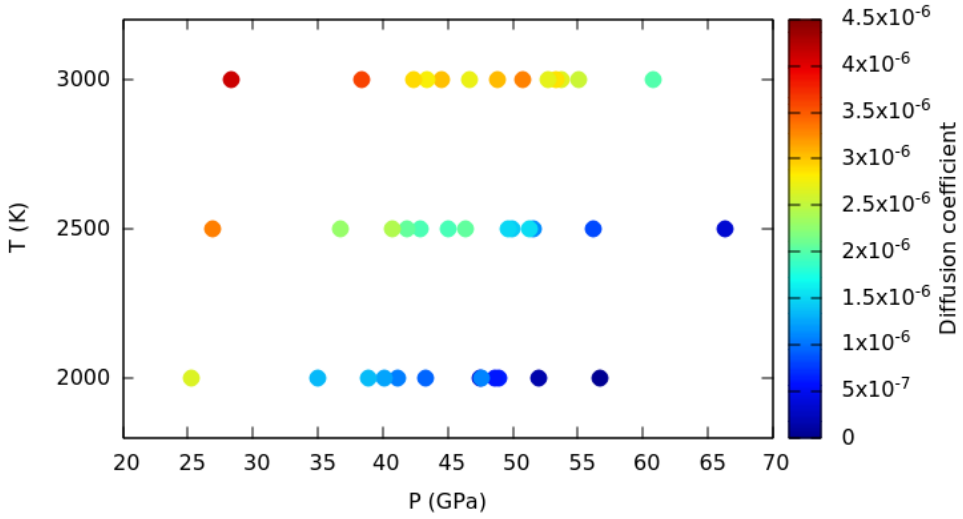


Figure 5.23: Diffusion coefficient of oxygen atoms as a function of pressure and temperature

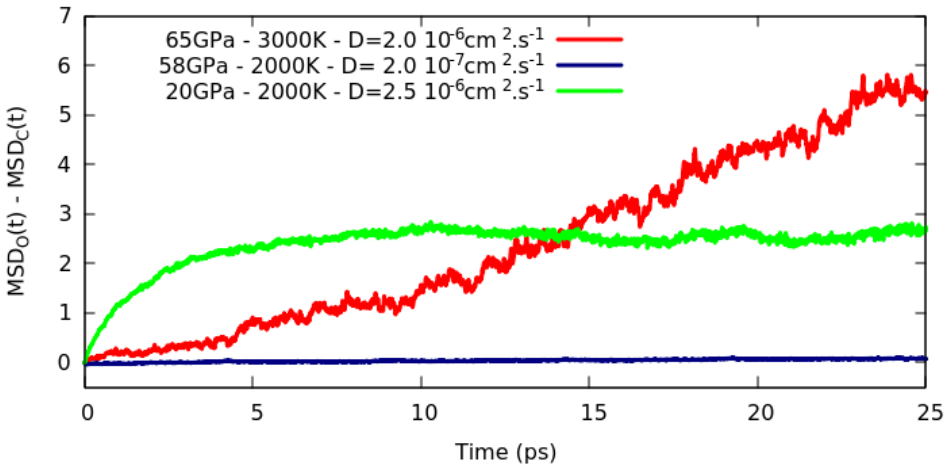


Figure 5.24: Difference of the mean square distance of carbon and oxygen atoms in the molecular, amorphous and polymeric liquid regimes

5.8 Chemical dynamics of the fluids

In order to analyze in more details the chemical reactions at play in the liquid, we used the *second layer local descriptor* (see 2.2). This allowed us to have a more comprehensive picture of the chemistry of the system than the coordination number of the high temperature polymeric liquid (65 GPa - 3000 K).

Although this description gives more details abot the general vicinity of an atom, the problem of the cut-off is amplified: indeed, the potential error on the cut-off will be present not only for the first but also for the second layer, and therefore, the error on the

bonds largely propages here and may create spurious results.

Assuming that this descriptor is valid, we can compute the fraction of carbon of all carbon state as in figure 5.25 and identify the states that are most present. In the case of figure 5.25, we see that the three most represented states are the one that are expected to be neutral, including the CO_2 molecule that is still largely present (21.5%).

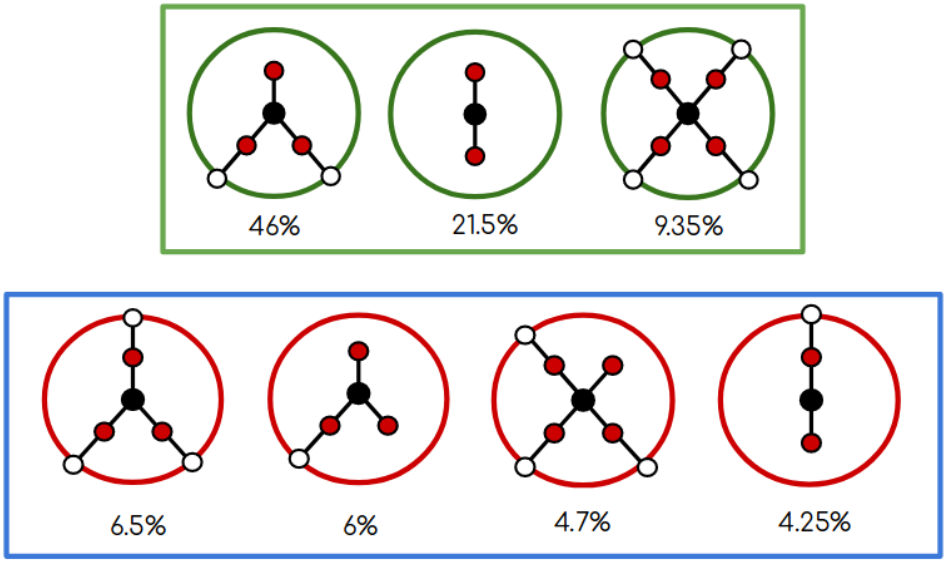


Figure 5.25: Second layer descriptors applied carbon atoms at 65 GPa and 3000 K, with their fraction of presence. Green circle indicate neutral states while red circles indicate charged states. States who were present for less than 0.1% of the whole simulation were discarded.

In order to analyze the chemical dynamics of the polymeric liquid, we needed to use a method that could somehow allow us to move past the issue of the flickering. One of the methods that presented to us was to implement a method similar than the one presented in [132, 60] mostly to study proteins: the Markov State Model approach.

However, here, instead of modelling the behavior of a whole protein using this approach, we aim at understanding the local evolution of atomic states. Therefore, instead of defining the states as different protein conformation, we used different atomic configurations (or states) instead. The approach, however is very similar: as we already have defined the states, we only need to compute the transition rates using (3.3). The main difference being that we do not compute the states by using clustering, but start from a given set that we postulate from chemical intuition. In order to check the validity of the Markov model, we can use (3.4), to test whether or not the Chapman-Kolmogorov equation holds.

As we can see in figure 5.26 the Chapman-Kolmogorov test seems indeed to mostly hold, at least in most cases. This implies that the system, at least at the atomic level, can be described as memoryless, and that we can use all the properties of Markov chains to

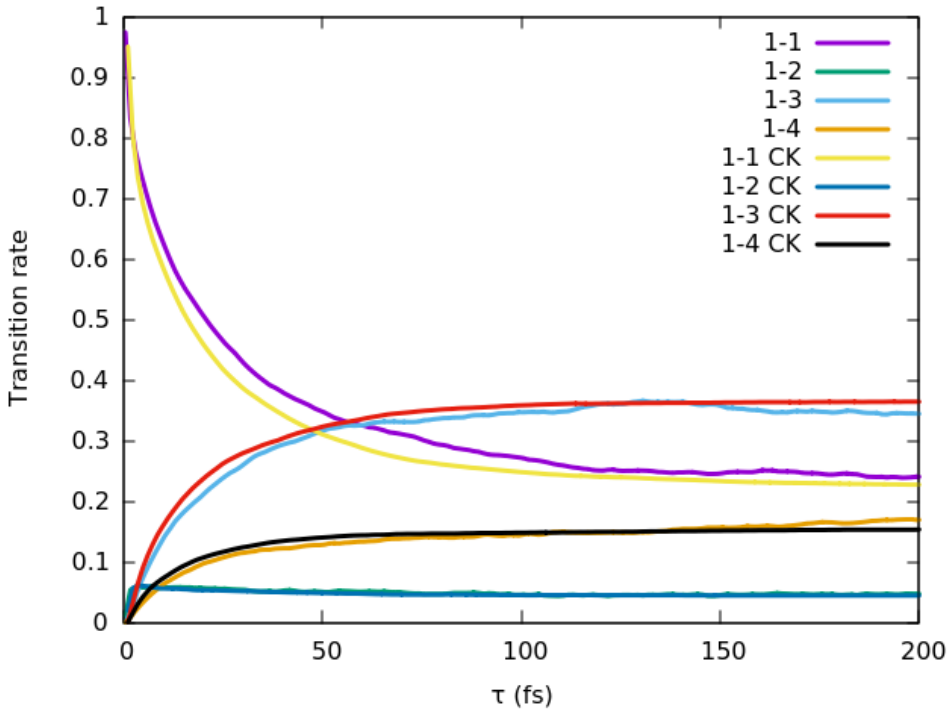


Figure 5.26: Chapman-Kolmogorov test for some of the transition rates of the carbon atoms. The CK in the legend indicates that the transition rate was calculated using the right hand side of 3.4.

compute various kinetic properties, for example the lifetime of the chains, which proved very difficult to do using standard methods, as the flickering of atoms around cut-off values strongly diminished their computed lifetimes (even after smoothing). Here, as the rates are composed using correlation functions, the flickering issue is mostly evacuated.

We do note however, that the Chapman-Kolmogorov equation does not hold all over the pressure range: it notably fails in the amorphous case and in some of the P-T simulations in the abrupt polymerization zone.

As this approach was developped relatively late we chose to focus less on the lifetimes of the molecules than on a (simpler) analysis of the transition rates between atomic states. Indeed, those transition rates allow us to gain important insight into the chemistry of the polymeric liquid.

In figure 5.27, we observe the isolated states and their transitions rates relative to the other states in the polymeric liquid at 3000 K and 65 GPa. We can first report that the charged states will have very short lifespans as their self-transition rates are low: between 7% and 18%, which implies that they are mostly intermediate states between the neutral states. Second, we see that the carbon dioxide molecules remains relatively stable, as it has the highest self-transition rates of the states (71%). Finally, the neutral threefold coordinated carbon is clearly the main chemical element of the liquid, as not only is it the most frequent (figure 5.25) it is also the state where most charged states will decay

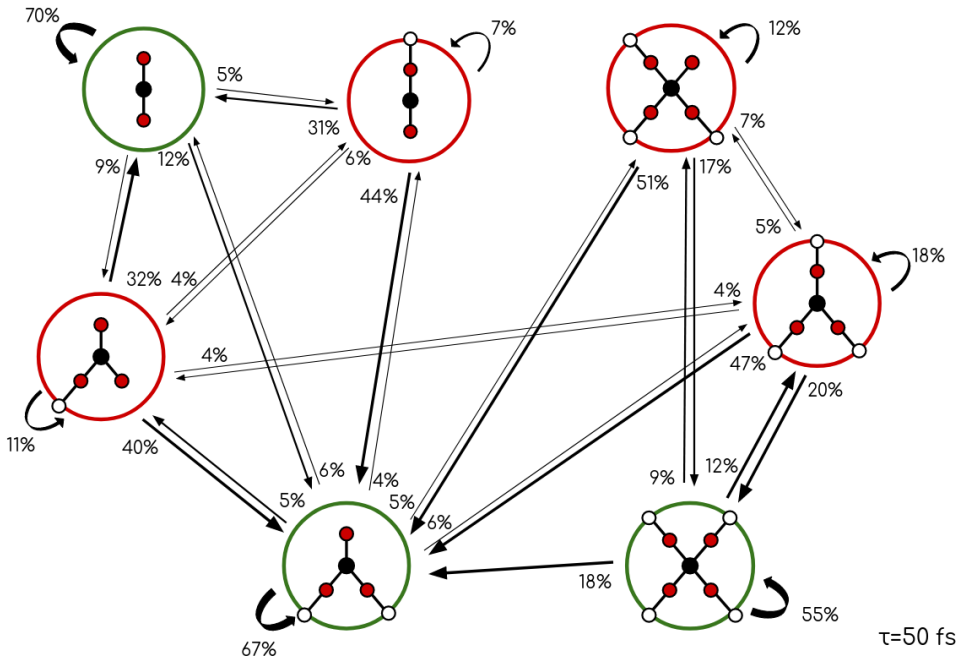


Figure 5.27: Dynamics of the carbon states at 65 GPa et 3000 K as seen by the Markov State Models. The percentages are the transition rates, computed using (3.3) with $\tau =$ 50fs. Configurations likely to bear a charge are indicated by red circle, while neutral states are in green circles. Transitions rates below 4% and states representing less than 1% of the total are not shown. The 180°arrows indicate the self-transition rates, the rates are written close to the emitting state.

to (all transition rates values from charged states to this state are on the order of 50%), and it also seems to allow the transition between the twofold coordinated states and the fourfold coordinated ones.

To conclude, the use of Markov State Models applied to atomic models in conjunction with the *second layer descriptor* seems to provide very interesting informations about the chemical behavior of the polymeric liquid. Although much work remains to fully exploit this new approach, those results are very much encouraging and can readily be applied over the pressure-temperature range where the system is markovian.

5.9 Conclusion

In this work we provide a full characterization of the behavior of carbon dioxide fluids in geological conditions. We put forward evidence of four distinctive fluid behavior: a standard molecular liquid, a reactive molecular liquid with formation of dimers and trimers, a highly reactive polymeric liquid and a amorphous-like liquid. Most of those behavior intersect the geotherm at one point or another, which suggest important consequences on the participation of carbon dioxide in the chemical activity and transport properties of the mantle. We show in figure 5.28 the final phase diagram that can be extrapolated

from our data.

We note that, due to the similarity between the amorphous phase and the one observed at low temperature through compression, it is likely that CO_2-V, a crystalline polymeric phase, may form instead in experiments, given that CO_2-V seems kinetically favored at high temperature in experiments compared to this phase.

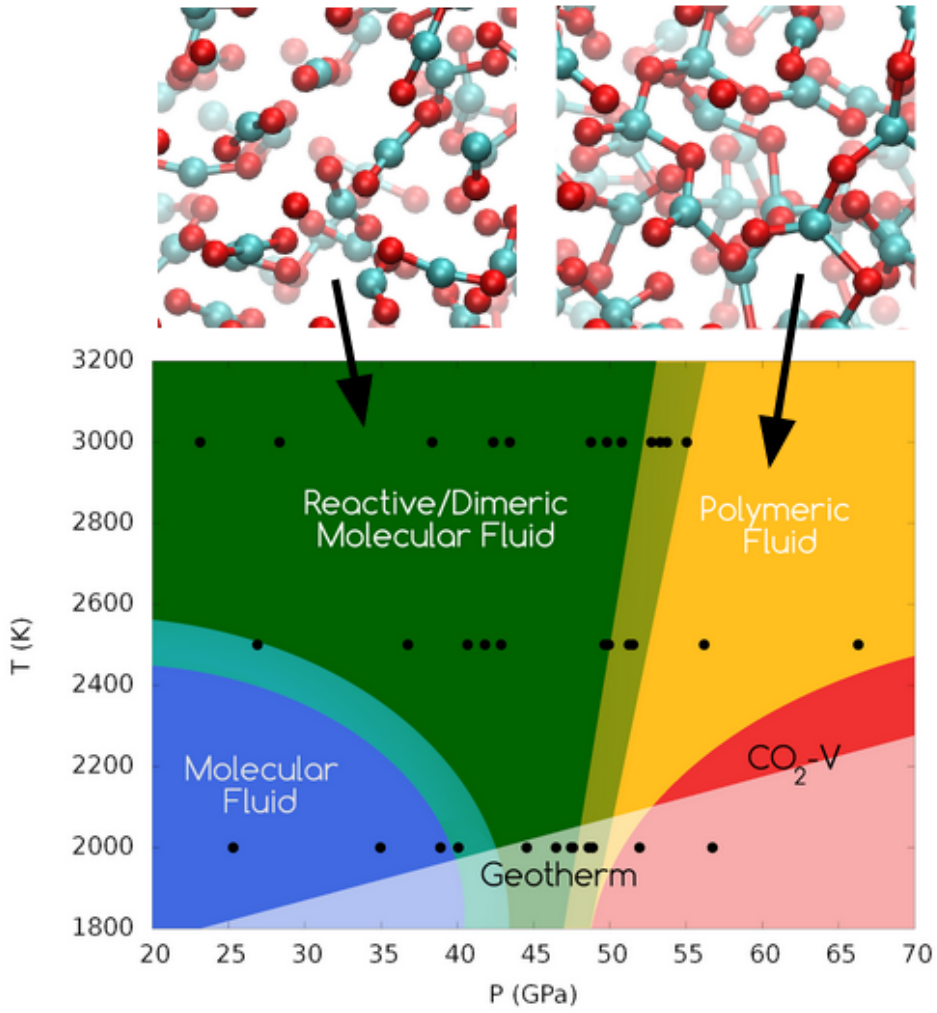


Figure 5.28: Propose phase diagram of CO_2 in geothermal conditions, with snapshot of simulation in the molecular (**top left**) and polymeric (**top right**) conditions. Black dots indicates the points where simulations were run. We use the frequency of C_2O_4 over 1.5% to mark the limit between molecular and reactive molecular liquid phases, the drop in C_2 carbons to mark the polymerization region, and the drop of the diffusion coefficient to mark the limit of Phase V.

The reactive molecular phase is of particular interest as it appears in conditions corresponding to the depth of Earth where carbon dioxide could form [4, 5], suggesting that the molecule may play an important role in the chemistry of the lower mantle.

Finally, we suggest that calculations with larger simulations box and in a less restrictive statistical ensemble (NPT) and larger simulations boxes may be necessary to fully study the free energy barrier between the liquid(s) phases and the polymeric ones. However, our calculations remain the most comprehensie study in this P-T region of high pressure carbon dioxide fluids and represented a significant computational cost.

Future works may leverage the amount of data generated during this work in order to construct efficient force fields that could be used to study more extensively the free energy landscape associated with the transformation between the molecular liquid and both the polymeric liquid and CO_2-V, which would be of high interest for geology.

Chapter 6

Crystallines phases

Introduction

As mentionned in the introduction, the molecular phases of carbon dioxide have been extensively studied both experimentally and theoretically [13, 17, 18, 20, 21, 22, 23, 65]. If the structures of the various phases are nowadays more or less consensual, the transitions mechanisms between those phases remain obscure.

In this small chapter we first show the results of our work on the transformations of the molecular crystal phases of carbon dioxide under high pressure. First we carried out a study of the various phase using *ab initio*, then we focus on the transition between the various molecular phases.

Although most of the phases were known before this work, we use the AIRSS method (see 4.3.1) to search the stable structure of carbon dioxide in the 0-80 GPa range. The objective was twofold: first to test the effectiveness of the method on a material that polymerizes at high pressure but it was also the occasion of a transfert of knowledge between an PhD student about to defend (Adrien Mafety) and a future PhD student (the author), so that the team did not lose this experience.

We then present the begining of our work on the phase transition of carbon dioxide. We used classical molecular dynamics based on a forced field that proved successful in previous works [51]. We use the metadynamics[133] method in order to explore the free energy landscape (see 4.1.1), using as collective variable the the Path Collective Variable (cf. 2.1.3) with the Permutation Invariant Vector (see 2.1.2) as internal descriptor.

6.1 AIRSS applied to CO_2

We started our work on the crystalline structure by using the AIRSS method (see 4.3.1) in order to check whether the method allowed to recover the known crystalline structures of carbon dioxide as shown in figure 6.1. The objective of this work was to recover as much as possible of those phases using a moderate computational cost, as it was mostly

a training exercice to learn the method.

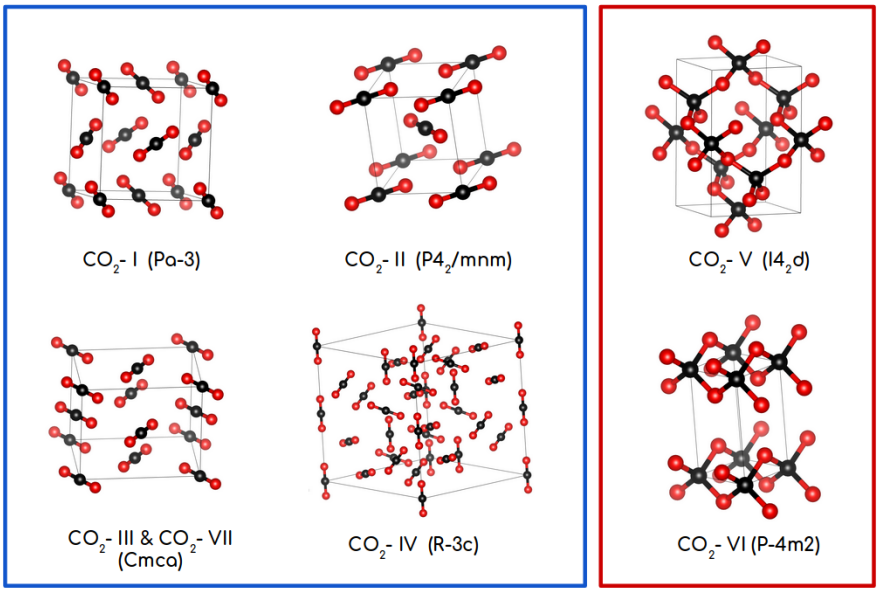


Figure 6.1: Crystalline phases of high pressure carbon dioxide

We launched AIRSS searches at three different pressures: 1, 5, 30 and 80 GPa, using the procedure in (4.3.1) to recover both molecular and polymeric crystal phases of carbon dioxide. For each pressure, we used two types of boxes containg either 2 or 4 CO_2 units. We did not include larger search boxes due to the potential high computation cost and therefore could not hope to find CO_2-IV whose convetional cell contains 24 CO_2 units. Each search was given 24h on 48 CPU cores to run, and generated between 300 and 1800 structures dependending on the pressure and number of CO_2 units.

The relaxations for the search of structure were carried out using *ab initio* calculations at the DFT level of theory with the CASTEP [158] code and Vanderbilt [78] pseudopotentials using 500 Ry for the cut-off for the kinetic energy of the electronic density. We then recovered the 10 most stable structures, after removing duplicates (using findsym [151]) and further relaxed the ten remaining structures using CASTEP [158].

At 0GPa and 5GPa we found the two structures reported for phase II ($P4_2/mnm$ and Pnnm) along with phase III (Cmca) and we did not find any non-molecular phase. Phase I (Pa-3) was not found in the original search but we were able to recover it by restricting the search to cubic cell.

At 30GPa we found a mix of polymeric and molecular phase, with crystalline polymeric phase CO2-V ($I4_2d$) as the most stable phase overall. At 80 GPa we found phase CO_2-V as the most stable phase and layered P-4m2 - one of the candidate structure for CO_2-VI - as a metastable candidate with close enthalpy. Although an alternative structure for phase VI, a layered structure ($P4_2/nmc$) - identical to P-4m2 except for the fac that the orientation of the tetrahedra are rotated from one layer to the next by 90° along

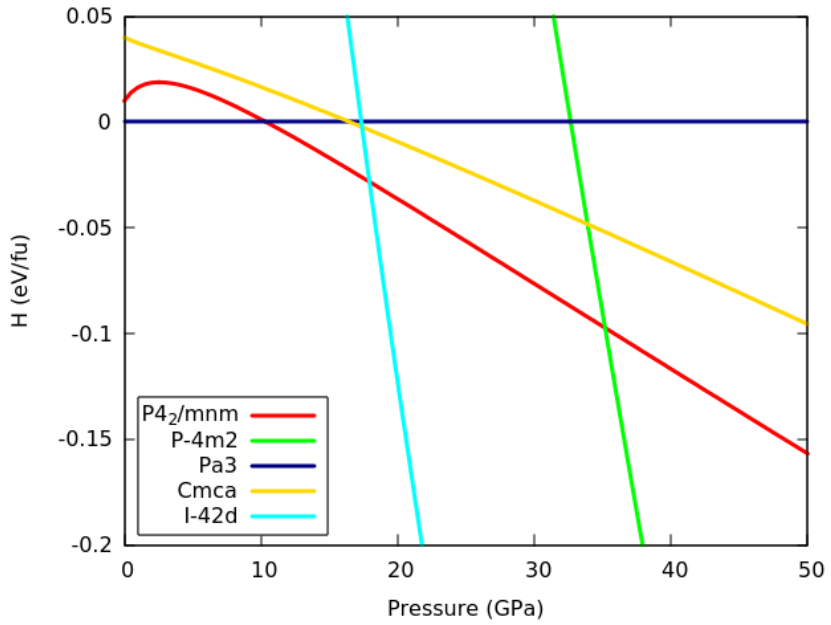


Figure 6.2: Enthalpy (by formula unit and relative to CO_2-I) of all CO_2 found with AIRSS, focusing on the molecular structure.

the axis perpendicular to the layer axis - has been predicted as more stable than the P-4m2 [34, 36]. we used the P-4m2 as reference to illustrate phase CO_2as both structures where proposed for phase VI and we could not recover the $P4_2/nmc$ layered structure with AIRSS. Interestingly in the theoretical work [34], P-4m2 formed from both phase II and III while $P4_2/nmc$ formed from III and IV, while using metadynamics [35], only P_4m2 was found upon compression of phase II. It is therefore likely that both phases could form, depending on the themodynamical path. We indicate that both structure have a very similar IR spectrum [36], which imply that only X-ray diffraction experiments would allow to differentiate which of the two structures is formed.

All the structures corresponding to reported phases were then relaxed between 1 and 50GPa[1] in order to get the 0 K theoretical phase diagram of high pressure carbon dioxide (figure 6.2) using quantum espresso[178].

Between 0 and 10 GPAa, as expected, phase I is the most stable structure, however we found that CO_2-II is more stable than phase III although in experiments, CO_2-I forms phase III under compression instead of phase I. Here we find that CO_2 become than phase I around 10GPa while phase III becomes energetically favored around 18 GPa only. The discrepancy can partially be ascribed to the martensitic nature of the transition between phae I and III, which is therefore kinetically favored to the I-II transition and therefore explains that CO_2-III forms preferentially from phase Ixs.

We also found that polymeric phase V is more stable than all molecular crystal phase

[1]We relaxed the structures at every GPa between 1 and 10 GPa, then every 10 GPa between 10 and 50 GPa

as early as 18 GPa, although the experimental phase transition is observed at 40 GPa above 1000K and around 60 GPa at ambiant temperature (figure 11). We expect that this is due to the large free energy barrier between the molecular and polymeric phases, which make it difficult for the system to polymerize. This is likely as phase V can be recovered at pressures as low as down to 1-2 GPa [32, 48], showing the importance of metastability effects in the transition from molecular to polymeric (and vice versa) system in CO_2.

In conclusion, we showed that AIRSS is able to recover most of the stable structures known for CO_2 at a moderate computational cost. We also were able to recompute the 0K phase diagram of carbon dioxide between 0 and 50 GPa for the found structures. Although those results were not new, and similar phase diagram had already been obtained, the results still confirm the ability of AIRSS which even with limited computation time is able to recover most of the stable structures of carbon dioxide.

6.2 Transitions between molecular phases

6.2.1 Context

Recently, Gimondi et al. (2017)[51] used well tempered metadynamics[139] and analysis in order to study the I-III phase transition, with classical molecular dynamics and rigid CO_2 molecules. Using the λ collective variables introduced in the method section, they were able to show clearly defined free energy wells for both phases, but the commitor analysis showed that this 2D projection was insufficient to understand the completely characterize the transition. However, The addition of the anisotropy, as defined by the ratio of the longest to the shortest cell length allowed the identification of a reliable transition pathway candidate, and to get a quantified value for the I-III phase transition. This work also found that packing faults CO_2-I like structures (I_def) may be more stable than pristine phase I, especially around the transition pressure. This study found a I-III transition, which is much less than the reported experimental pressure (10-12GPa, but with significant hysteresis depending on the transition path), but in accordance to previous study done using static CO_2 molecules.

6.2.1.1 Computational details

We used classical molecular dynamics we used the same the three-site Trasnferable Potential for Phase Equilibria (TraPPE) as [51]: the CO_2 molecule was taken as rigid, with the C-O bond fixed at 1.16Å, and two types of interactions were used: Coulomb and Van der Waals using the same parameters as in [51]. We used boxes with 864 CO_2 molecules, that is 2592 atoms in total, as a compromise between the precision of the calculations and their costs.

As mentioned above we used the Path Collective Variable with Permutation Invariant Vector (2.1.2) as underlying descriptor. We used only the oxygen atoms in the construction of the collective variable in order to limit the cost of calculation, the carbon atoms

giving only redundant informationsabout the structure. We used a switching function of the same shape as presented in (2.1.1) and chose the parameters (n=5, m=10, r_0=5Å) after checking the swithching function with regard to the pair correlation function. This choice of parameters results in a very smooth switching function able to take into account the second shell of molecules around the oxygen atom (figure 6.3).

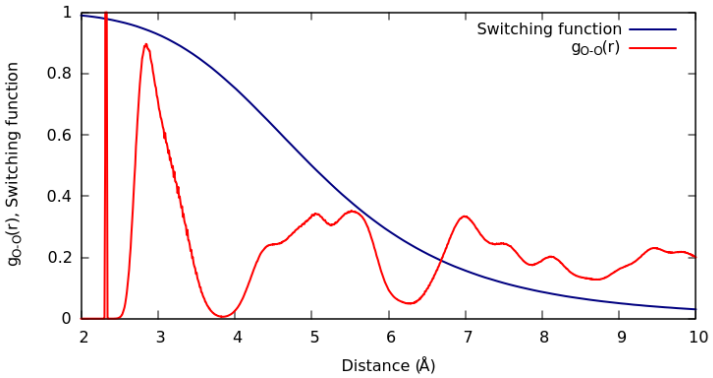


Figure 6.3: Switching function and O-O pair correlation function

All calculations were carried out using GROMACS[100], in the NPT ensemble using the velocity-rescale thermostat, with a time parameter fixed at 1ps. We also used a Berendsen [102] barostat with a time parameter at 10ps and a timestep of 0.5fs. The LINCS [99] algorithm to maintain the rigidity of the CO_2 molecule.

6.2.2 Results

In order to prepare the simulations cell, we equilibrated all phases using 5 ns equilibration simulations every 5 GPa between 1 GPa and 35 GPa both, at 300 K and 600 K still in the NPT ensemble and starting from either the structures obtained in the AIRSS search or [159]. Using the relaxed structure, we computed the PIV distances between all the phases, and we builty a topological map in PIV space of the various phases (figure **??**).

The idea of the topological map is to project all structures as point in a plan so that all the distances between the points in the plane structures is equal to the actual distances between the corresponding phases. Interestingly we see some ressemblances between the topological map and the phase diagram of carbon dioxide: we at least expect that the distances between phases should somehow mirror their ability to transform into one another in experiments. We do find encouraging results in figure 6.4 with phase I be far from both phases II and IV, with III in the middle of the path between them. This is nicely relatable to the fact that phases III and VII which share the same structure act as intermediary in between phase I and phases II and IV. We do find that phase I and the molecular fluid seem to be far from each other, however, this is likely due to the fact that the chosen liquid phase was equilibrated at 10GPa, where it is most closely related t phases IV and VII[6].

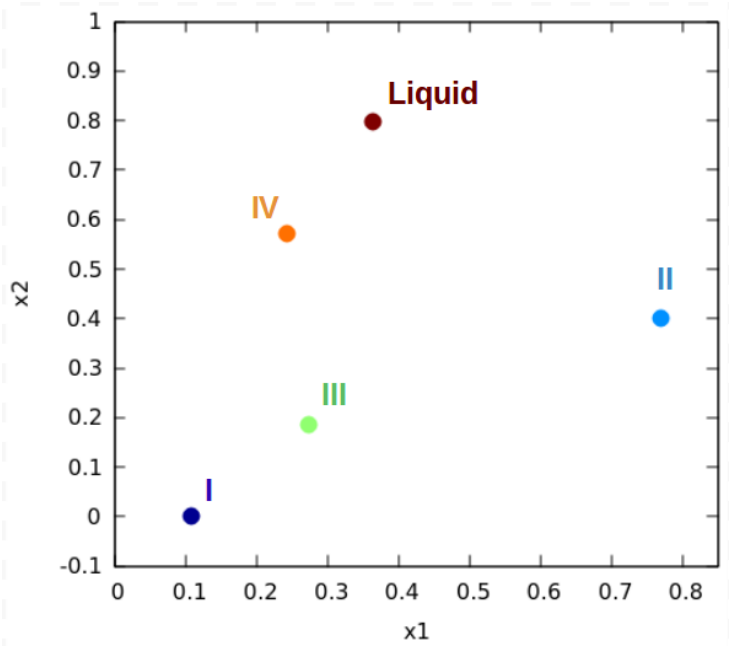


Figure 6.4: Topological map of the molecular crystals phases in PIV space (distances are normalized with respect to the largest one).

We then focused on the transition between phases I and III, which was already explored by [51] with the same force field at 350K. We expected this transformation to be relatively easy to observe as it requires relatively small topological changes. In order to do that, we started a molecular dynamics simulation accelerated by metadynamics using the equilibrated phase I at 5GPa and 300K as initial position. We used the equilibrated phase I and III at the same conditions as references. We used gaussians of 5 kJ.mol^{-1} and width of 0.01 in the S variable and 0.1 on the Z variable which were added every 500 steps. We used a 0.005fs as in the equilibrations. The temperature and pressure were maintained at 300 K and 5 GPa, using the same algorithm as in the equilibration as well.

Using this set up, we were able to observe the transition from phase I to phase III in a relatively small amount of computation time (less than 1ns of simulation), and the return from phase III to phase I was oberved after a much longer time ($\sim$ 10ns). Using the coarse results of the exploration of the landscape by the metadynamics, we managed to reconstruct the free energy landscape (figure 6.5). Comparing the height of the barrier and the difference in energy between the two phase we observe the same results as [51]: we find a barrier of $\sim$1200 kJ.mol^{-1} between phase I and III, and a difference in free energy around 800kJ.mol^{-1}. However, we find the transition between I and III in a relatively straightforward way by simple elongation of the cell along the c-axis and a rotation of CO_2 molecules into a plane and found no evidence of the deffected phase I that was found by [51], although we cannot exclude that it was not due to the limite simulation time that we used for this simulation.

Although we were able to repeat the experience at the same temperature and pressure as [51], we found that at 400 K and above, the simulation would invariably lead the trans-

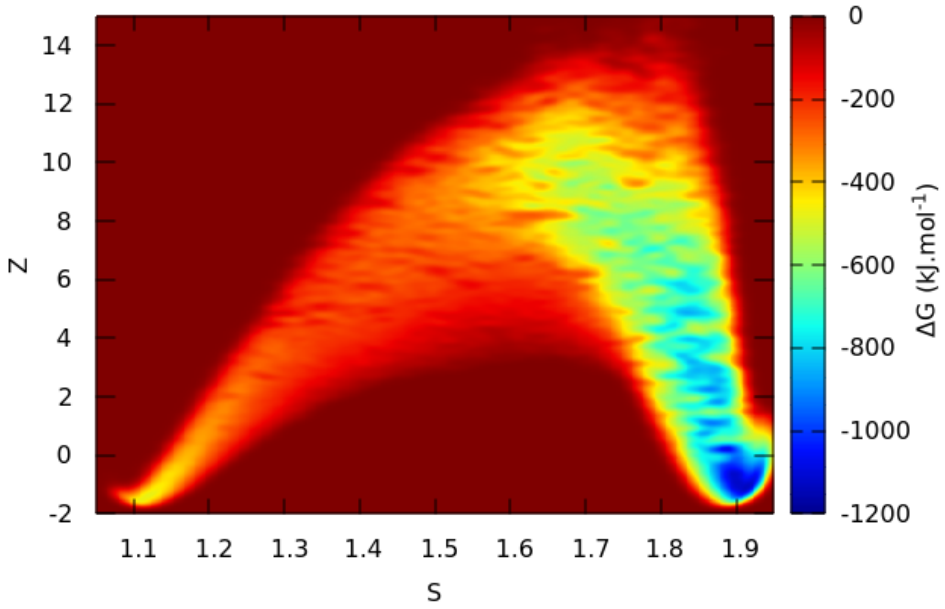


Figure 6.5: Free energy landscape of the I-II energy landscape. Phase I is in S=1.1,Z=0 and phase III in S=1.9,Z=0.

form into the molecular liquid phase, even using repulsive bias to avoid this situation. We found similar sitations when studying the transitions between phase I and phaes II and IV: in both cases, we managed to observe the transition to phase III at 300K, but not the transition from phase I to the target phases, nor did we observe a III-II or III-IV transition. At higher temperature, all transformation lead to the liquid phase, and the same was observed at all temperature when studying the II-IV transition.

Overall it does seem that the force field is probably the limiting factor. Indeed, although it is relatively accurate to describe the structures of the various phases[6], it seems to be much less accurate when computing the difference in energy between phases: indeed, we found that it predicted CO_2-II and IV less stable than CO_2 over the whole 0-30 GPa range at 0K, but also more stable than CO_2-I up to 0.5 GPa, which is more problematic. However we cannot exclude that some issues with the metadynamics simulation or collective variable definition may not also be at fault.

6.2.3 Conclusion

Although we only managed to reproduce the results by [51] in so far as we observed the transition at 300K and 5GPa between phase I and III, it is enough to show that the combined use of PathCV and PIV is able to efficiently explore the transitions between bulk structures in CO_2, as it was for water [113]. In general our results are in agreement with those of [51], although it seems that we were not able to find the deffected phases

proposed in this study.

Although those results are encouraging, more work can be done to further exploit this method, even with potential limitations of the force field. The FES in the I-III transition should be more rigorously explored using umbrella sampling for example, and it should also be worthwhile to compute the height of the barrier between phase I and III as a function of the pressure.

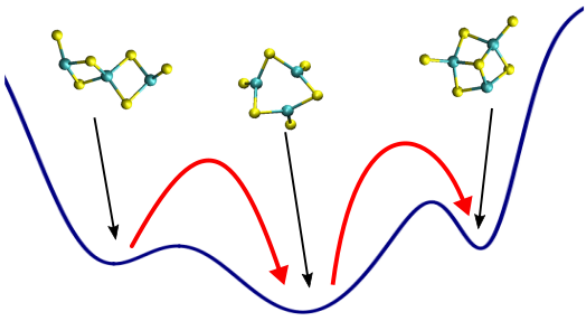
Finally a methodical analysis of the force field accuracy seems to be needed to go forward. In this case, as in the case of the liquid-liquid transition.

Part III

Satellite projects

Chapter 7

Unsupervised explorations of configurational space applied to MoS2 clusters



This project was a follow-up of the work accomplished during a master student project, in collaboration with a fellow graduate student ***Sofiance Schaack****. The project's objective shifted progressively during its progress and it resulted in a paper untitled* **"Unsupervised computer exploration of MoS_2 nanoclusters: structures, energetics, and electronic properties"** *that is currently submitted to The Journal of Physical Chemistry C.*

7.1 Introduction

In recent years molybdenum-sulfide materials have been widely studied for a variety of applications, ranging from hydrodesulfurization to transistors. In particular those material can form interesting nanostructures with potential exotic properties such as inorganic fullerenes and nanoplatelets that have either been synthesized or are highly sought after. This is a strong motivation to study for the formation mechanism of Mo_nS_2n nanostructures, which may help gain insight into the formation of the larger forms, as well as provide information about the relationship between the configurations and their properties.

Mo_nS_{2n} clusters where regularly studied in closely with W_nO_{2n} clusters due to their strong chemical similarities. Both of those type of cluster have been the focused of both experimental [160, 161] and theoretical [162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 173, 174, 175] studies in the past, focusing on the energy stability of the form and through the investigation of several stochiometries "magic clusters" (clusters forms that are much more stable than their counterpart) where found. Interestingly, their structure seemed to suggest that those configurations tended to favor over-saturation of sulfur and monolayer like forms as opposed to 3D arrangements[176] like core-shell structures.

In most previous theoretical studies used methods based on enumeration of chemically intuitive geometries, using heuristic arguments or similar structures as that of the bulk. This kind of approach risks missing structure as it lacks objectivity and transferability to other system with different atomic species, for example. As mentioned in (4.3), the need for an effective and unbiased search method to search for stable configurations is well appreciated in the condensed matter community[146, 61] in general and in the nanostructure field in particular [177]. In this respect, two recent studies used evolutionary algorithms to explore the configuration space of MoS_2 [175] (although the search was supplemented in this case by human-provided structures *ex post*) and WS_2 clusters [174].

In order to supplement this information we present in this project a new exploration method and apply it to small MoS_2 cluster of three different sizes (Mo_2S_4,Mo_3S_6,Mo_4S_8). Our alternative searching method uses a combination of *ab initio* molecular dynamics, enhanced sampling methods (metadynamics) and topological collective variable (SPRINT) that captures the topological environment of the system along with clustering technique (using Daura's Algorithm (3.2.2)).

7.2 Exploration methodology

The method that we used in this project to identify stable clusters can be divided in 6 steps:

- 1 : Sample the configuration space with metadynamics
- 2 : Identify candidate structures
- 3 : Initial relaxation of candidate structures
- 4 : Remove duplicate structures
- 5 : Relax candidate structures
- 6 : Computate of electronic/vibrational/other properties (optional)

The first step is generally the most expensive: configuration spaces tend to be high dimension landscapes which makes them complicated to explore, and even more so to sample. Metadynamics can overcome this issue, provided that the time needed to overcome the various barriers between the configurations is not too high. This tend not to

be true for bulk materials but it tends to be the case for small clusters, where the computation cost of a single step is small even in an *ab initio* setting. Another advantage of metadynamics is that, contrary to many other methods, space is explored progressively which in turns implies that the the mechanisms of the transitions may be recovered from the trajectory, or at the very least from subsequent simulations which will benefit from the data acquired by metadynamics.

It is also important to mention that metadynamics require efficient collective variables in order to be productive. Such collective variables should properly captures the differences between various configurations that can be explored by the system. The technique also needs to be well calibrated, as seen in 4.1.1.

The second step consists in choosing proper candidates that might relax into stable configurations so as to recover all possible stable configurations (or as many as possible) and to have the fewer amount of duplicates. For this work, we have essentially used two different methods that we will detail here.

At the beginning of the project we decided to use a method solely based on the results of metadynamics. Here the identification of candidate forms relies on the variations of the bias potential:, as metadynamics fills up free energy wells, we expect that the bias "felt" by the system increases as the observed configuration have already been explored. However, when the system falls into another unexplored well, the bias should significantly drop[1]. Those drop therefore can be used as a good signal that a new configuration is being explored. It should be noted that the height of the drop does not accurately represent the height of the energy barrier between two successive states.

Another effective signal that we used to determine stable configurations is that while the system stays stuck in a given potential well, the collective variables should do not significantly, as the system's configuration does not evolve much (assuming that the collective variables are effective). Therefore, the relative temporal stability of the collective variables may be used as another sign of stability of a configuration while large variations of the collective variables are associated with transition states. In practise we used both this input and that of the drop of the bias potential to determine candidate structures.

Although this method seem appealing and the selection criteria may seem pragmatic, the application of the method is however, quite costly in human time. The method reqires that the whole trajectory be manually analyzed, which is fast when only a few wells are found, however, in practise we found that there could be a huge number of drops in a single metadynamics simulation, which made it impractical. As we cold not find a way to automatize this analysis we therefore looked for an alternative.

The second method takes another wholly different approach to the previous one: the idea here is less to identify potential stable structures, but to provide an efficient tessella-

[1] In some cases, especially after long simulations, wells may be filled without being explored due to the spread of the gaussians

tion of the explored space loosely based on the local explored point density. In our case, we used Daura's clustering (cf 3.2.2) algorithm in order to do so. As the metadynamics will result in higher density of points in the region where deep free energy wells were observed, corresponding to stable structures, the tessellation should roughly be able to follow the landscape and the cluster center can then be used as candidate structure for relaxation. In order to do so, the d_c parameter needed to be adjusted so that to limit the number of duplicates (and therefore useless relaxations) but also not large enough as to lose potential candidates. The advantage of this approach is that it requires relatively little intervention from the user while still being effective.

Once candidate structures have been chosen, we used a two step geometric relaxation. The purpose of this two step optimization is to save time as the first relaxation is coarse and threfore fast, and aims at roughly finding the nearest well, while the second step is here to fully equilibrate the remaining structures and is more expensive and more precise.

This two step process, although not necessary, has the advantage to cut down on computation time as structures that may originally look different may fall inside the same well and be caught as duplicate at an earlier point. It is also possible to automatize the search for duplicates by discarding all structures which distances in collective variable space and difference in energy are both below user-defined cut-offs. Once the final relaxation step is completed, it is finally possible to compute the properties of the system.

7.3 Computation details

In order to explore the configuration space *ab initio* molecular dynamics calculations accelerated with metadynamics [133], at the DFT level of theory. We ran several simulations for each of the three distinct Mo_nS_{2n} configurations ($n = 2$, $n = 3$, $n = 4$) of interest. We ran at least five different simulation per cluster size, aiming for simulations longer than 100ps. Originally, we used the Quantum Espresso code [178], but as the scaling was not satisfying for the *ab initio* molecular dynamics/metadynamics part, we switched to the CPMD[157] code. In this chapter we only present the results obtained with CPMD. Metadynamics algorithm was implemented through the PLUMED[112] plugin, using in both cases version 1.3.

The simulations were done in the Born-Oppenheimer approximation[68] with a timestep of 1fs, the cut-off for the wavefunction was put at 120Ry and we used a convergence criterion of 10^{-5}Ry for the self-consistent cycles. We used Perdew-Burke-Ernwerhof (PBE)[179] functionnals employing Goedecker-Teter-Hutter[180, 181] pseudopotentials to describe core electrons. Although several box sizes were tried, our results were obtained with cubic cells of 12Å side. Temperature was maintained around 600K using a Nose-Hoover thermostat with a cut-off frequency of 1500cm^{-1}.

The total trajectory lengths for each cluster size are: 250ps for Mo_2S_4, 900ps for Mo_3S_6 and 1200ps for Mo_4S_8, amounting to a total simulation time of 2.35ns. Note that several

calculations did not reach the 100ps objective because the calculations stopped because the calculations stoppped (in general the system either dissociated or the configuration obtained was too high in energy to be solved by the software). The resulting trajectories were kept and a new calculation was in this case re-launched from the final configuration (or in some cases, a configuration obtained a few steps prior), with all bias removed.

For the metadynamics part of the calculations we used the SPRINT[55] collective variables (cf 2.1.1) using the following switching function:

$$\sigma(d_{i,j}) = \frac{1 - \left(\frac{d_{ij}}{d_0}\right)^n}{1 - \left(\frac{d_{ij}}{d_0}\right)^m} \tag{7.1}$$

The parameters of the switching function were estimated using the pair correlation function for each cluster size. This pair correlation function was obtained through a 2ps *ab initio* molecular dynamics simulation without bias.

For all cluster size we chose the following values: $d_0 = 4.5$Å, n=8, m=24, so that the network would account at least for the second shell neighbors. We fixed the height of the gaussian potential bias as $H = 0.04$eV, the width as $\sigma = 0.7$, the gaussians were deposited every 10fs. Those parameters were chosen using small test runs, insuring a good compromise between exploration speed and stability of the simulation.

All the obtained trajectories were concatenated for each system size to obtain a final meta-trajectories (lengths are given above). Daura's algorithm was then used on each of the meta-trajectory in order to get a tesselation of the collective variable space using the piv_clustering code[54]. Due to memory limitation inherent to the implementation, we divided each meta-trajectory into smaller size samples of 250ps. The choice of the d_c parameters was made so that a few hundred of cluster centers emerged per cluster size. In order to account for the increase of dimensionality of the CV space with size, the values of d_c were increased for larger system sizes: $d_c = 0.3$ for Mo_2S_4, $d_c = 0.6$ for Mo_3S_6 and $d_c = 1.2$ for Mo_4S_8. The cluster centers were then used as potential stable structures and relaxed.

The candidates structures were first in a two step method as described above, and the duplicates where filtered out of the remaining structures based on structural and energy similarity. A second, more stringent relaxation step was then done, in order to fully equilibrate all structures, so as to be able to compute electronic or vibrationnal properties if need be.

Both relaxations were run with the Quantum Espresso code [178], using *ab initio* calculations at the DFT level of theory and Perdew-Burke-Ernwerhof (PBE)[179] functionnals and Rappe-Rabe-Kaxiras-Joannopoulos ultrasoft pseudopotentials[182] in cubic box of 15 Å side. For the first relaxation, we used a cut-off of 60Ry on the kinetic energy for the

wavefunction and 720Ry, while we used a cut-off of 120Ry for the wavefunction and 1440 for the density for the secondary relaxation. In both cases we used a convergence criterion of 10^{-3} a.u. on the forces to obtain the final geometries. Spin polarization effects were taken into effect only in the last relaxation, in order to compute magnetization of the clusters. In order to compute the binding energy, single SCF were run for a single atom of Mo and S in the same conditions as the secondary relaxation, in order to compute the energy of isolated Mo and S atoms in a box of 15Å.

7.4 Results and discussion

Following this method we identified 109 stable clusters: 14 structures for Mo_2S_4, 27 for Mo_3S_6 and 68 for Mo_4S_8. For each of the clusters we computed the binding energy (BE), Highest Occupied Molecular Orbital -Lowest Unoccupied Molecular Orbital energy gap (HOMO-LUMO gap) and magnetization (μ_B). In the context of this work, the binding energy (BE) was defined as follow:

$$BE = \frac{nE(Mo) + 2nE(S) - E(Mo_nS_{2n})}{3n} \tag{7.2}$$

Where $E(Mo)$ and $E(S)$ are the energy of the isolated Mo and S atoms, respectively, and $E(Mo_nS_{2n})$ is the energy of the relaxed cluster. This is strictly equivalent to study the stability of the cluster with regard to a fully dissociated system, and is most widely value used to compare the energies of clusters in the literature as it allows to compare the energies of configurations of different sizes.

We used this metric to sort the clusters energetically in each cluster size so that for each size n, we identify them as $n.rank_n$. For example the 5^{th} most stable structures of Mo_3S_6 size is referred as *3.5*. For each size, the eight most stable structures and their properties (BE, HOMO-LUMO gap, and magnetization μ_B) are presented in figure 7.2.

The observations of the most stable structures the results found in previous works [175]: for $n = 3$ and $n = 4$, the most stable candidate has the form of the 1T monolayer phase of MoS_2. We further observe a clear tendency to form platelets instead of core-shell structures.

We also find several structures similar to those predicted using an evolutionary algorithm on WS_2 clusters[174] and we are able to recover all previously predicted structures of the litterature [175]. We signal on figure 7.2 the structures that were previously found using colored circles: green for structures found in studies where MoS_2 structures were the focus and orange in the case of WS_2 clusters. Blue circles indicate the presence of a bulk-like structure.

In addition to the eight most stable structures we also present in figure 7.2 several motifs that are shared by many of the stable configurations (indicated by colored squares in the same figure):

- an Mo centered tetrahedron (blue square)
- a five member ring with three Mo atoms (two of them bonded), two S atoms and an capping S atom (violet square)
- an 3 or 4 member ring of Mo atoms (green square)

We also showcase several interesting, albeit less stable, forms in figure 7.2 that stand out from the others by their symmetry or specific shapes. Notably we observe:

- two very symmetrical structures (*2.10* and *3.10*)
- a ring-like structure (*2.13*)
- the begining of a 1D structure (*3.10*)
- an S-S-S chain on an otherwise stable structure (*4.20*)
- a platelet-like structure that does not match the proposed structure of the monolayer phases 1H nor 1T (*4.21*)

Another interesting aspect is the evolution of the energy as a function of the size of the forms as presented in figure 7.1. The main aspect that is visible is that globally, the larger structures tend to be more stable than their smaller counterpart. We also note that the most stable structure of Mo_3S_6 is separated by a relatively large BE gap (0.1eV) to the next most stable cluster (the same difference is of 0.028 for Mo_2S_4 and 0.017 for Mo_4S_8). The same type of gaps between successive structures is however observed in Mo_2S_4.

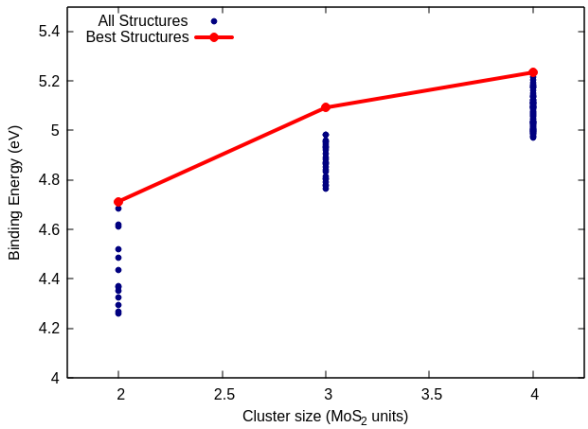


Figure 7.1: Evolution of the binding energy (in eV) of the most stable configuration with increasing size (in MoS_2 units) in red, along with the binding energies of all stable structures (blue).

It is not generally straightforward to compare our results with published results as various functionals and/or optimization scheme were used. We do observe however that

our ranking matches those presented in [175]. The single exception occurs for two structure whose BE difference is very small (0.02eV) and is thus more likely due to the use of the differences in computation scheme mentioned above.

We observe the same kind of results while comparing the HOMO-LUMO gap and magnetization of all clusters. We find a good general agreement with other publications when determining whether or not structures are magnetic or non-magnetic, but we find different values of magnetization. The same kind of agreement is found for the HOMO-LUMO gaps, where we have a general agreement in the order of magnitude, but where large variations are observed in the difference between our results and those of other studies. Both of those properties being highly dependant on the description of the electronic structure those discrepancies are expected due to the fact that different *ab initio* schemes have been used.

We note that DFT corresponds here to a good compromise between precision and computational cost for the exploration and geometry optimization, to uncover the stable structures. However, if one is interested in the electronic properties, it is likely that a much more demanding computation method (such as RPA, Coupled Clusters, etc...) should be used.

7.5 Conclusion

Although the method that we propose here is very promising and yielded interesting results, it is i,portant to point out some limitations and areas where further work may be necessary. The main cost of the method comes from the metadynamic. This means that the main computation limitation is the ability to carry out *ab initio* molecular dynamics simulations long enough to allows if not all then most of the configuration space to be explored. The scaling of the method is therefore very much that of DFT and this means that it is highly dependant on both the system's size (in terms of electrons) and that of the cell. It is also necessary to point out that the size of the configuration space to explore should increase with the size of the system, although the exact scaling is not easily computable.

Despite the aforementioned limitation, in this project we have showcased an effective exploration method combining *ab initio* calculations, enhanced sampling and clustering algorithms . We demonstrated the effectiveness of this method on Mo_nS_2n nanoclusters: we were able to not only recover all previously found structures but also to find new ones without making any guess as to the structures of the stable forms. We find excellent agreement with those of a similar study using evolutionary algorithm [174].

In terms of the cluster themselves, further work may be done by analyzing the various transitions between stable structures, especially close-lying ones that are near the minimum of energy. More involved calculations may also be performed in order to get more precise values for the electronic properties such as the HOMO-LUMO gap and magneti-

zation.

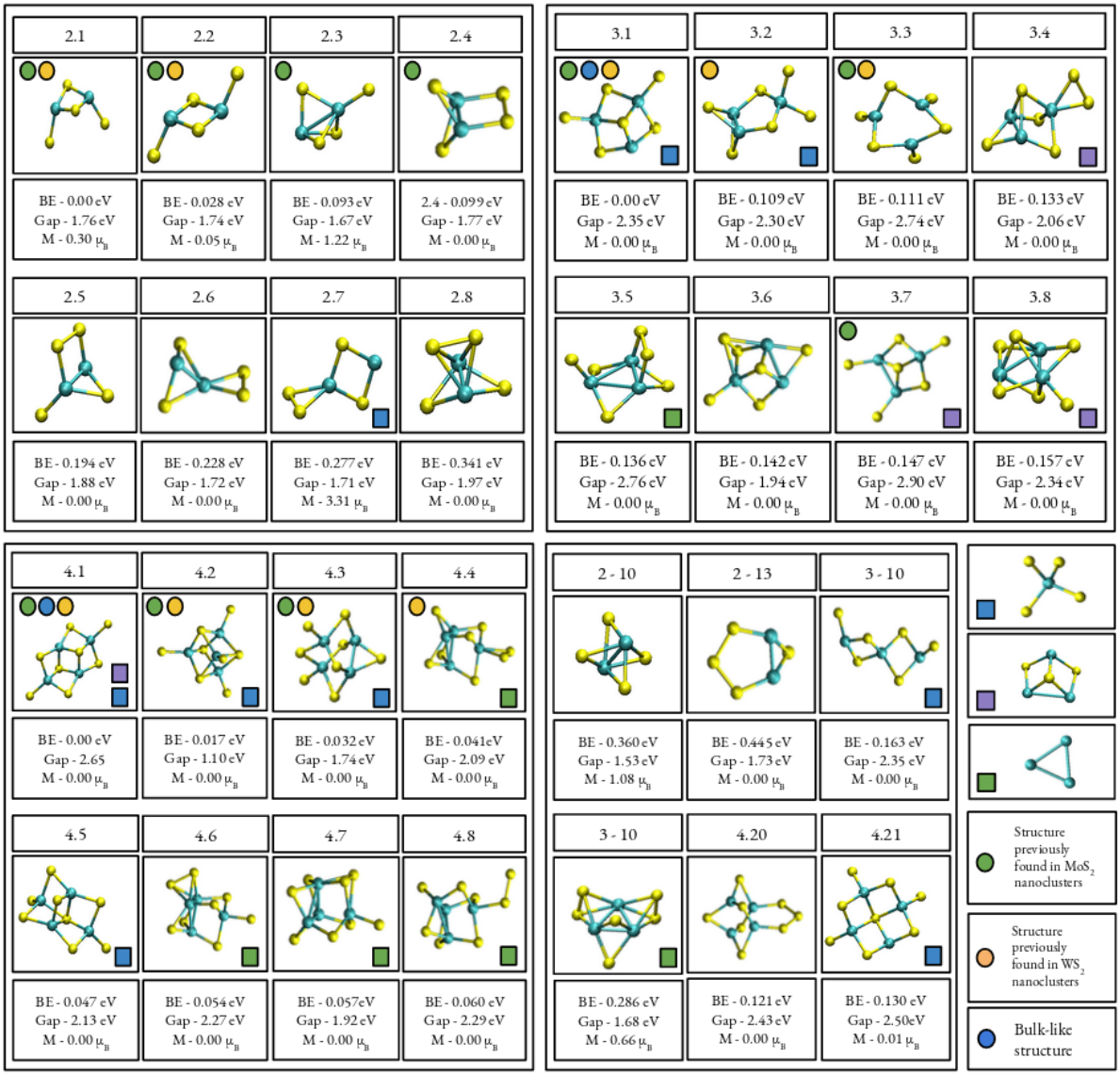


Figure 7.2: Most stable configurations and associated binding energy, HOMO-LUMO gap and magnetization (n=2:top-left; n=3:top-right and n=4:bottom-left), and less stable but nevertheless interesting structures found for all sizes (bottom center, along with their physical properties). The bottom-right part lists the recurrent motifs and the color/-type code indicating the presence of the motifs and/or whether or not the corresponding structure was already found in a previous work.

Chapter 8

Boosting *Ab Initio* Random Searching of Structures

This project was conducted by ***Gabriele Mogni*** *and supervised by* ***F. Pietrucci****. The author contributed to the project as advisor mostly on the use of the AIRSS method, ran some of the calculations and by proposing a convergence criterion (see below) for the exploration method. This projects has lead to a paper that is currently being written and tentatively untitled:* **Understanding the topology of energy landscapes explored by crystal structure prediction algorithms**.

8.1 Context

As mentioned in 4.3.1, there is currently a lot of scientific effort invested in the elaboration of structure search algorithm. The *Ab initio* Random Structure Searching (AIRSS) [61] is one of the most famous method that was created for this very purpose. This method (which has been described in 4.3.1), although very efficient, suffer from a number of drawbacks:

- **Poor scaling:** the method scales relatively poorly with the size of the system, as it relies on the ability to perform a large number of geometric relaxation with *ab initio* calculation. As this type of calculations scale with the number of electrons, the cost of the method makes it impracticality for systems larger than a few atoms per unit cell, especially if the atoms have large atomic number. This is amplified by the dimensional scaling of the configuration space with the number of atoms, implying that not only are the unit calculations more expensive but one needs of much more of them to recover a large number of stable configuration as the system grows.
- **Absence of stopping criterion:** the method does not provide a clear criterion to stop the search. Indeed, the proposed method proposes to stop when the user deems that a large enough number of structures have been relaxed.

The aim of the project was therefore to improve on both of the limitation, using the case of high pressure silicon phases (around 1GPa) as a test case. This test system pre-

sented the advantage to be relatively simple in terms of number of atoms (one type of atoms, few atoms necessary) and has been extensively studied, which allows us to have some reference points [61, **?**, **?**].

8.2 Preliminary analysis

In order to study the effectiveness of AIRSS, we started by generating 20.0000 structures containing 4 atoms of Si in a cell whose volume was fixed either around 60Å or 80Å, in order to have some enough data to compute statistics about the method. The structures were relaxed using the quantum espresso software[178] system).We used PBE functionnals[72] and Ultrasoft Vanderbilt pseudopotentials[183], the cut-off for the kinetic energy of the wavefunction was put at 150eV and k-point spacing of $0.03Å^{-1}$. All structures were then optimized at a simulated pressure of 1GPa.

An overview of the results of the relaxation is presented in figure 8.1 where we show the distribution of the structure as a function of the volumes of the relaxed structures as a function of their enthalpy. The range of volumes for the stable phases is large as indicated in previous work[61], some phases having very similar volumes :simple hexagonal, beta-tin and Imma around $60Å^3$; while other have widely different volumes: cubic diamond around $80Å^3$ and I4/mmm around $85Å^3$ (figure 8.1). The range of energies is also very large with a difference of 0.6 eV between the most stable structure (cubic-diamond) and that of the highest energy structure associated with a target phase (Imma) (figure 8.1). We also observe that the Imma phase has an interesting distribution of the volumes ranging from $80Å^3$ for the most energetic one to just under $60Å^3$ for the most stable one (figure 8.1.

We first tried to see if there was underlying information about the underlying FES in the topology of the random structures. In order to do so, we computed the probability that a given structure relaxing into a given Si_4 polymorph is neighbor to another structure relaxing to the same polymorph within a given distance in PIV (cf 2.1.2) space for 5000 randomly chosen structures out of the original 20000 set of randomized structures. In order to have proper values of distances, we used 3x3x3 supercells of the orignal randomized structures, and we repeated the analysis for 4 of the 5 polymorphs of interest.

For the computation of PIV we used a switching function of the form:

$$\sigma(d_{i,j}) = \frac{1 - \left(\frac{d_{i,j}}{d_0}\right)^n}{1 - \left(\frac{d_{i,j}}{d_0}\right)^m} \tag{8.1}$$

where $d_{i,j}$ is the distance between atoms i and j, d_0, n and m are free parameters allowing to tweak the range of distances of interest. In our case, we chose $d_0 = 4Å$, $n = 4$ and $m = 8$, corresponding to a switching function capturing the at least the second nearest neighbor layer of atoms for each atoms.

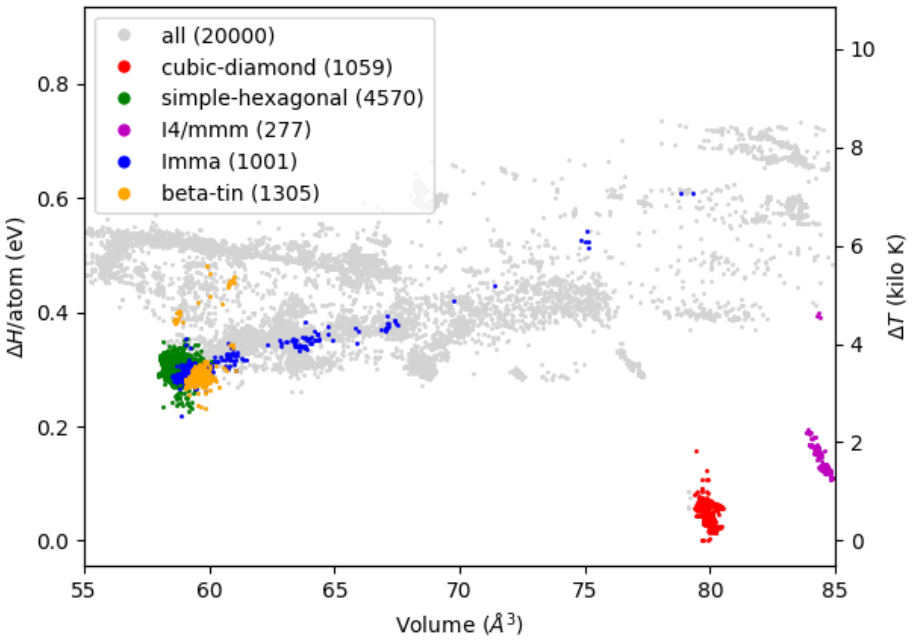


Figure 8.1: Enthalpy per atom in eV as a function of the volume for each relaxed structures found using the AIRSS method. The corresponding values of the energy in temperature are shown on the left y-axis in K. The number in the parenthesis relates to the number of structures that relaxed into the target phases.

The results show that when no relaxation is done, there is in fact very little information that is contained in the topology of the system. However, we see that even with short relaxations, we start to see important correlation between the structures, although this varies widely from one structure to the next.

The main point to take from this results is that the structural properties alone of the initia configuration does not contain any clue about the underlying energy landscape. However, we are only using half of the available information: indeed, whenever we generate a given configuration we can compute its energy as well, which when combined with the topological distances between phase may yield more results. We tested this hypothesis by computing the distance PIV of every structure to the lowest energy ones as a function of the difference in energy between them (figure 8.3).

In figure 8.3 we see the progressive sampling of structures, from 5k structures in the top panel to 20k in the bottom one, and we see that some sort of tail emerges, pointing towards the origin. This tail indicates that at 20K structures, the energy well of the cubic-diamond phase (that is, the structure associated with the lowest energy in the sampling) starts to be sampled with many different structures (figure The existence of this tail may be used as a visual criterion that the configuration space is correctly sampled, and so to provide a stopping criterion assessing that the landscape has been explored.

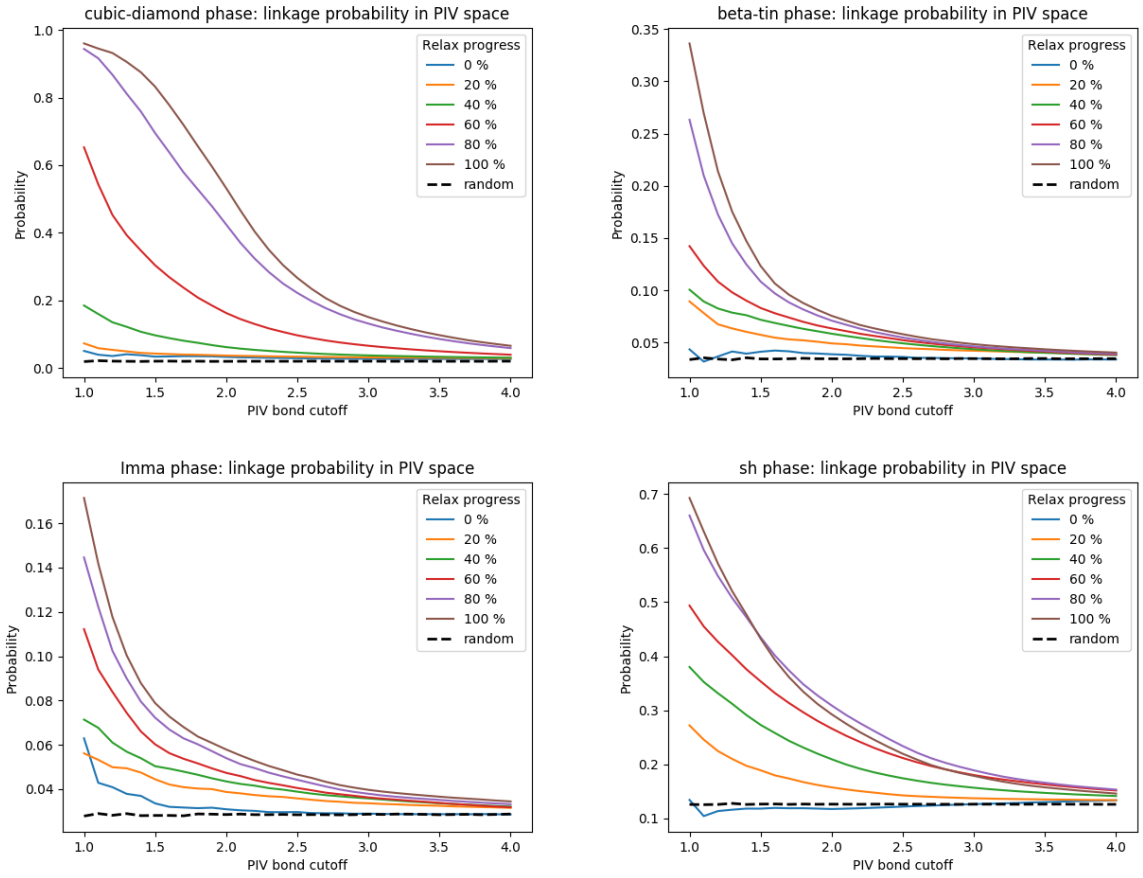


Figure 8.2: Fraction of structures within a given PIV radius of a structure that relax into the target phase. The progress of the relax is measured on a scale of the maximum of 50 step of relaxation that were allowed (20% means 10 relax step).

What is more, this results seems to indicate that by simply evaluating the energy of the initial structure and the distances between them in PIV space, it might be possible to drastically reduce the number of relaxation necessary to find the relevant (meta)stable structures, therefore adressing the first limitation of AIRSS. Indeed, using only local evaluation of the energy, it seems to be possible to find at least some stable structure and to sample the landscape. As those evaluations are exptected to be cheaper than geometry optimization, this may cut computational cost significantly.

The second limitation of AIRSS was the stopping criterion indicating that the exploration had converged. In principle it is impossible to know when all minimum have been found, however it may be possible to find a weaker criterion assessing that the method has arrived to the point where it finds only structures that have already been found. The tail in figure 8.3 could be used as such a criterion but it is mostly a visual one and can be complicated to use.

In order to overcome this, we propose to use the variation of the Density Peak Clustering (cf 3.2.3) The idea here is to use the energy (ρ) as the density in the original algorithm,

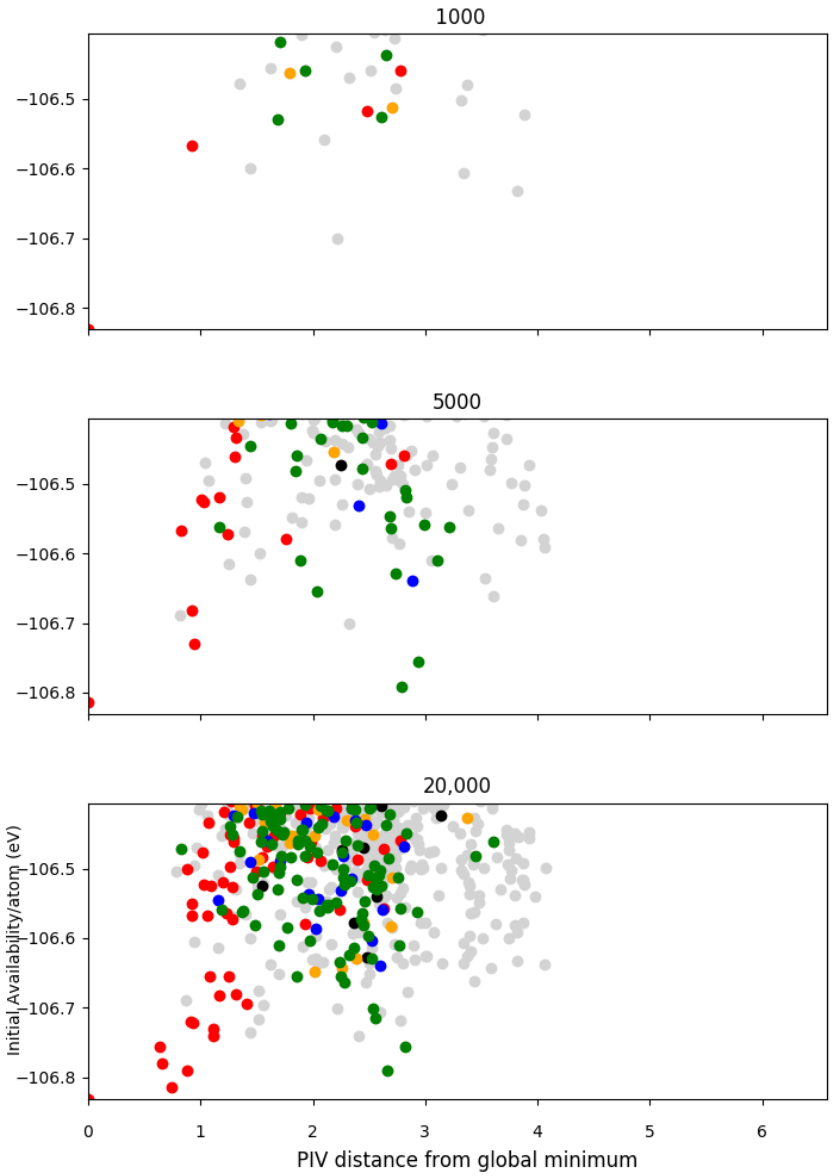


Figure 8.3: PIV distance to the most stable structure as a function of the difference of energy with the most stable structure, for three sets of random structures: 1000 structures (top), 5000 structures (center), 20.000 structures (bottom). The colors corresponds to the final structure that the random structure relaxes to: green for simple hexagonal, red for cubic diamond , blue for Imma , orange for beta-tin , black for $I4_mmm$. Grey dots correspond to points that relaxes to phases other than the five main reported phases. The range of energy has been cut to contain only structure within 0.5eV from the most stable one.

and distance to the closest point of higher density (δ) by the PIV distance to the closest point of lower energy.

In doing so, and following the same procedure as described in (3.2.3) to find extremum

in sampled landscape, it should be possible to identify the stable candidate structures. However, and more to the point, one can imagine that the decision diagram of this method should progressively converge as the more and more structures are added corresponding to sampling of the configuration space.

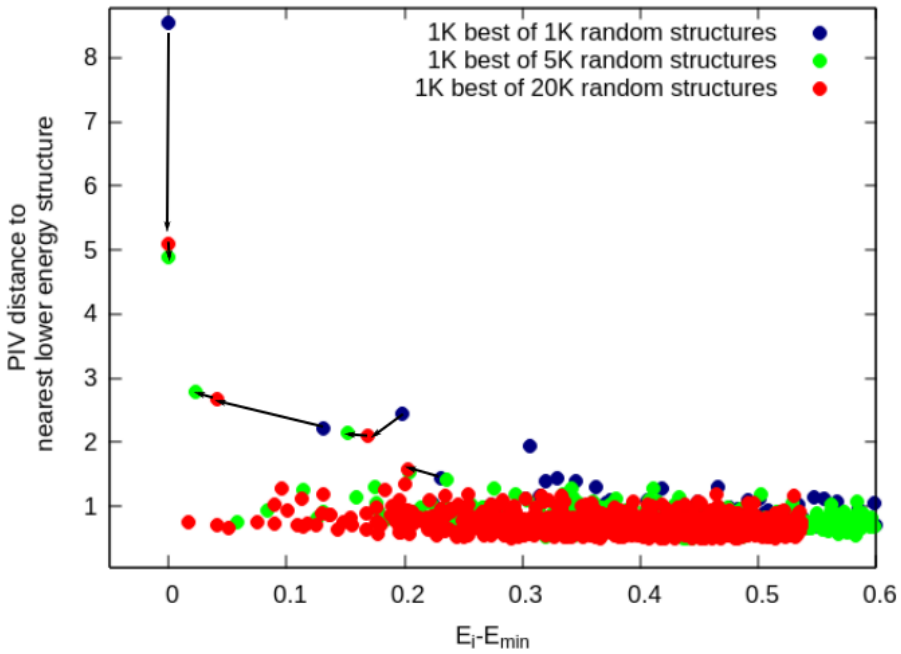


Figure 8.4: Decision diagram related to the progressive sampling of the configurations space of Si crystals: for each of the 1000 best structures for each of the set of 1000,5000 and 20000 structures, the difference of enthalpy with regard to most stable structure for each stable structure is shown as a function of the PIV distance to the nearest structure (in terms of PIV space) of lower enthalpy

In this regard, we show in figure 8.4 that we find that the decision diagram inded seems to converge. This is a good sign that this diagram may be used as a stopping criterion. Although not all stable phases were found systematically found by the Density Peak Clustering method, the convergence criterion still worked even in those cases.

In conclusion, we therefore have of two methods that could adress of limitations of the AIRSS and we therefore can propose a modified version of the AIRSS that we will call AIRSS-Boost.

8.3 AIRSS-Boost methodology

Using the conclusion of the preliminary analysis, we can combine the elements of sampling and the decision diagram-based convergence criterion to implement a modified AIRSS algorithm that can be summed up as follows:

- Generate a large amount of initial configurations ($\sim$ at least several hundreds) and evaluate their enthalpy.
- Compute the decision diagram of the set
- Repeat step 1 and 2, until the position of the cluster centers in 2 converges.
- Remove potential duplicate structures, keep only the structures within a given energy range of the most stable one
- Relax of the remaining structures
- Computation of the material properties of interest

From that point, the main question that remains is the potential gain in computational cost of the method compared to standard AIRSS. Indeed, it is still possible that scanning properly the phase diagram may prove to costly, especially in high dimension space, while random structure relaxation may require less simulation to converge.

8.4 Comparative results

In order to compare the efficiency of both methods, we computed the probability for each method to have found 4 of the target structures (one of the stable configuration was ignored as it was found to be very difficult to find) for a given amount of computing hours. The resulting graphs are shown in figure 8.5 where the both methods are compared for three different number of initial configurations for AIRSS-boost: 5000,10000 and 20000.

The probability here is computed by evaluating the average frequency that a given AIRSS/AIRSS-boost method would have found all structures for the given amount of computing hours.

From figure 8.5 it seems that the methods start fairly even in terms of results for 5000 initial structures but that by increasing the number of initial structures in the sampling, the AIRSS-boost method gains a significant advantage over the standard AIRSS method, by a factor of 2 in terms of computing hours. An alternative AIRSS-boost methods, where the Daura algorithm was used in order to select the potential candidate structures is also used, but its results are at least equivalent, if not worse than that of the AIRSS-boost method.

8.5 Conclusion

Although most of the results that we obtain seem very encouraging, there is still much room for progress. First, the issue of the scaling of the method is not solved by the method, indeed, as the number of *ab initio* increases, the poor scaling of this kind of calculations for heavy atoms or in the cases where magnetism is present may still heavily

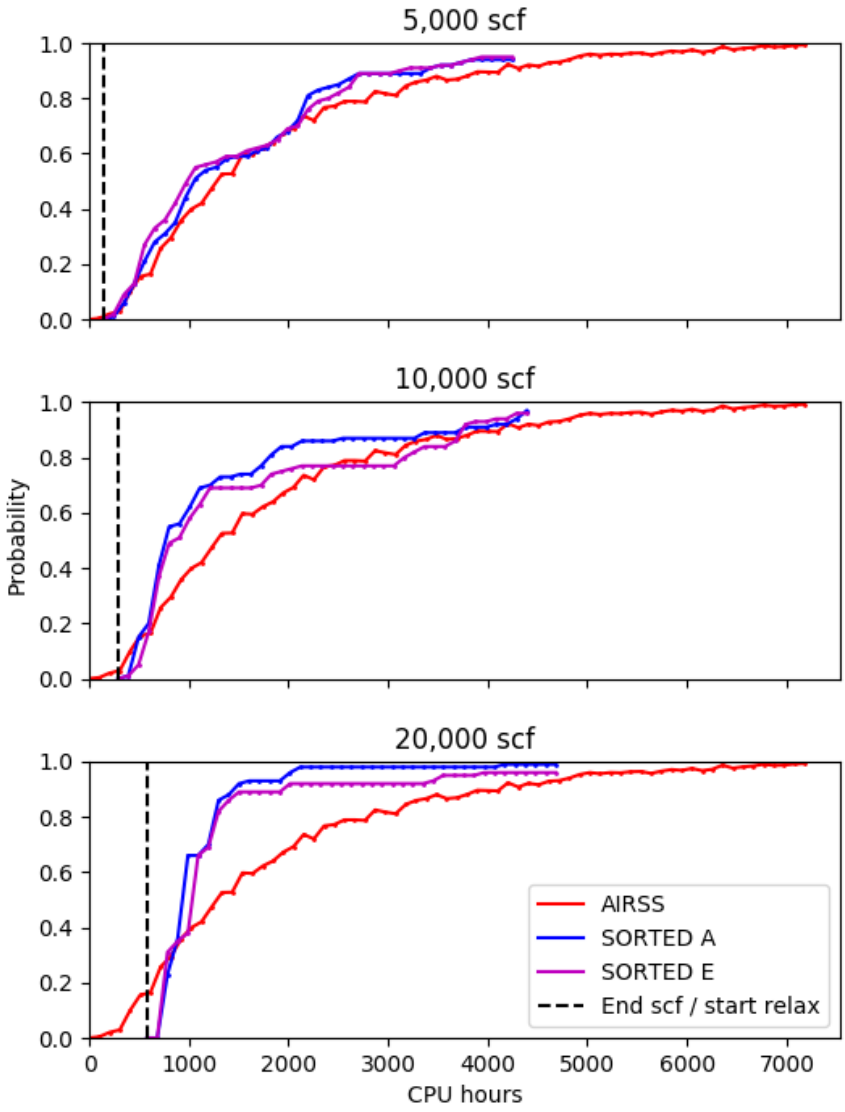


Figure 8.5: Probability that the four target structures are found as a function of the computing hours for AIRSS, AIRSS-boost (A) and AIRSS-boost with Daura Clustering (E) for three different amount of initial configurations. The initial cost of the scanning is indicated by dotted lines.

impact the cost of the method. Second, as mentioned before, this method may not be efficient in cases where the energy landscape contains a large amount of wells with small spatial extent, which would make them complicated to spot by the random sampling.

It is therefore important to test this method on a more complex (and/or heavy) system, and compare the results with the standard method of AIRSS. Regardless, the results

contained within this project shade new light on AIRSS and on its efficiency which should, if nothing else, at least contribute to a better understanding and use of this method. Further, a stopping criterion is established that can be used on either methods, proposing a solution to one of the two main issues reported at the beginning of the project.

Conclusion

This work focused on the transformation of carbon dioxide under extreme pressure (between 1 and 70GPa), corresponding to those of the lower mantle of Earth. We mostly focused on the progressive transformation of the molecular liquid into a polymeric liquid under geological conditions. We show that the polymerization of the molecular liquid results in two different polymeric fluid behaviors, one highly reactive, one amorphous-like. In addition, we show that the molecular liquid transforms progressively into a highly reactive liquid molecular liquid, that forms in conditions where carbon dioxide is likely to form in the lower mantle.

We also present two satellite projects focused on the creation or refinement of methods of search of structures, the first on nanoclusters of MoS_2, the second focusing on bulk material with the example of high pressure silicon. Both of those methods show great promise for future applications.

Apart from those applications, one the most important aspects of this work (in terms of time allocated to it) was the developpement of methods for analyzing the local environement of atoms. We introduced new methods in order to analyze the bonds between atoms using the Electron Localization Function, that prove extremely useful to provide an criterion based on the density to check wether or not bsome atoms are bonded. Those methods may also be viewed as potential starting ground for electronic-based collective variables for *ab initio* simulations: indeed, it seems feasible from this work to build a contact matrix containing ELF information rather than about distances between atoms. We also showed that clustering algorithm may be applied to specific local descriptor to divide them intro groups based on their local topology. Finally we point out that using both ELF and clustering may be another good starting point to build better analysis tool.

We also present a first succesfull application of the Markov State Model to the kinetics of atomic states, giving us important clues about the chemistry of the polymeric fluid. The results of this methods are very promising. Other applications of Markov Models that can be proposed from this study include the use of Hidden Markov State Models to also analyze the bonds in a system, using the large amount of statistics to guess the most likely sequence of bonding states corresponding to the observed distances.

Overall, we presented in this work the results of three years of work, focusing mainly on the transformation of carbon dioxide at high pressure and high temperature, but also on many satellite projects(some of which are presented in the appendices), which are

[1] all linked together by the conception of methods to analyze the behavior of complex molecular systems and the exploration of high dimensional landscapes associated with their configuration space.

[1] almost

Appendices

Appendix A

Polymeric Crystalline phase V

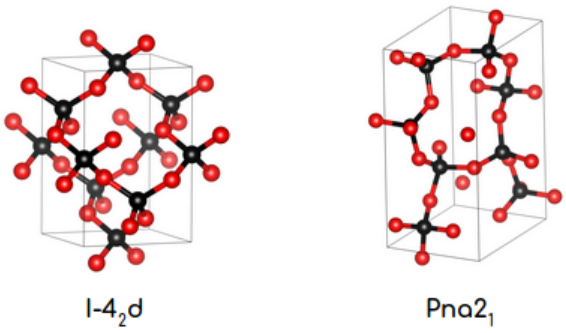


Figure A.1: Structures of phase I4$_2$d and Pna2$_1$

In 2016, Yong et al[48] used *ab initio* molecular dynamics accelerated by metadynamics in order to explore the landscape in the high pressure conditions where CO_2-V was found stable and propose alternative structure for phase V than the one currently accepted as best candidate [32]. In order to do so, they launched *ab initio* molecular dynamics calculations accelerated by metadynamics from the mechanically unstable structure originally proposed as a candidate for phase CO_2-V by Yoo et co-workers [25, 26], using a similar metadynamics algorithm as [59] which used the cell parameters as main collective variable.

From those calculation they found a new non-molecular and stable crystal phase, that they compared with the experimental X-ray diffraction results of CO_2-V. In order to put to the test the new structure as a candidate for CO_2-V we computed its raman spectrum, along with that of the commonly accepted I-4$_2$d structure. The configurations were first relaxed using the quantum expresso[178] code pw.x, and we then computed the raman spectrum from the resulting structures using the ph.x and dynmat.x codes of the same software.

Interestingly, the structure Pna2$_1$ would not automatically relax into the rapported structure unless we restricted the relaxation to orthorombic symmetries. Indeed, it would otherwise results in a slightly more stable structure featuring a slightly non-orthorombic

cell.

Once properly optimized at 40 and 50 GPa and we were able to compute the spectrum of both phases at those two different pressure. We report the results in figure A.2 for the 50GPa ca1se.

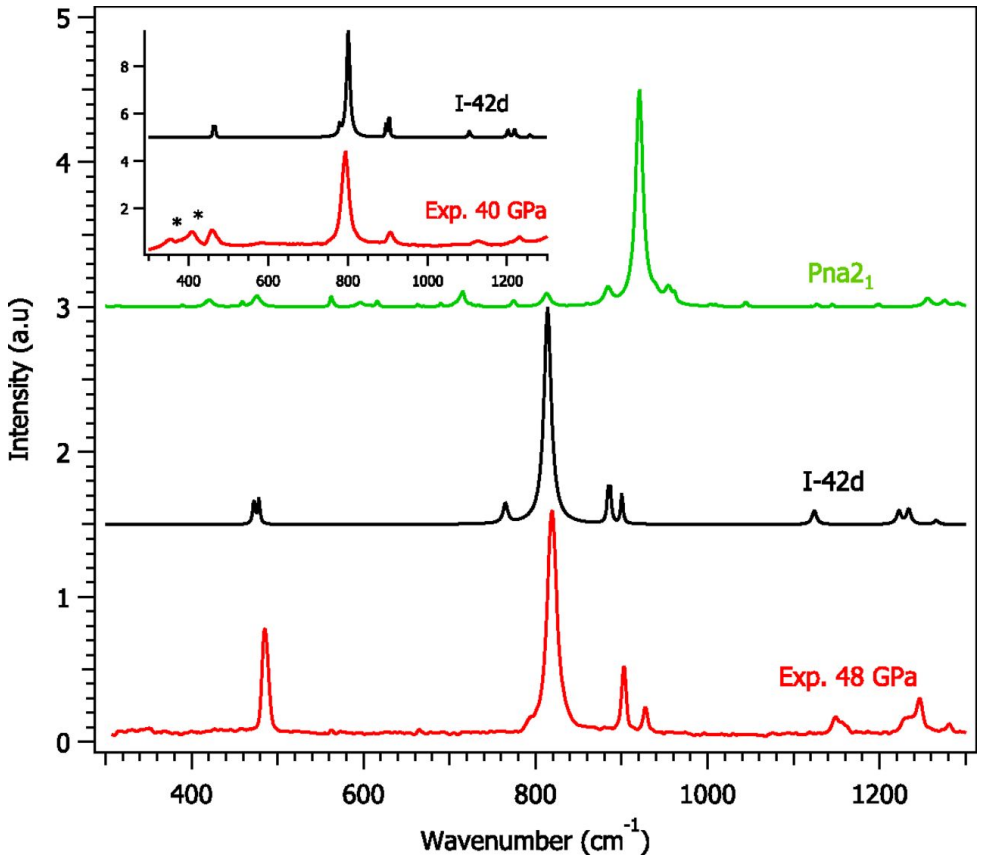


Figure A.2: Experimental spectrum from CO_2-V compared with theoretical predictions for structures I-4$_2$d and Pna2$_1$. Figure extracted from[184]

The resulting spectra seem to indicate that phase Pna2$_1$ does not faithfully reproduce the results, predicting many peaks that do not appear in the experimental spectrum, while I-4$_2$d seems to reproduce the experimental data. This piece of evidence, compounded with other remarks resulted in a comment[184] on the original paper [48] written by F. Datchi.

An answer was given to the comment[185], proposing a shifted spectrum for phase Pna2$_1$ (that still did not account for the experimental spectrum) and precised that the proposed phase Pna2$_1$ was not actually proposed as a candidate for CO_2-V but as a candidate for another extended stable phase for high pressure CO_2-V.

Appendix B

ELF in the middle: Application to Glycine solvated in H_2O

??

Here we present an small extension of our work similar to that presented in 5, in the analysis of the bonds in the system using the value at the middle of the distance between two atoms to determine whether or not they were bonded. We apply the same methodology on a system with a more atomic types, containg not only carbon and oxygen but also nitrogen and hydrogen.

We used data from a very short simulation of a glycine in water, computed by Andrea Perez-Villa during the course of her work in the team on pre-biotic chemistry. The software used was CPMD[157] both for the short trajectory and the ELF. The system contains an intermediate state to the formation of glycine from CH_3-NH_2 (methylamine), a CO_2 molecule and 80 water molecules.

The results of the analysis of O-H, C-O, C-N, N-O bonds are visible in Figure B.1. As the system does not evolve a lot during the very short trajectory, not much bonding change happends, however, there are still some point of notes.

On the O-H density, we can clearly differentiate between bonded hydrogen ($d <$ 1.2Å)and non-bonded hydrogen ($d >$ 2.5Å), and although the ELF still is able to indicate a clear difference between bonded (ELF > 0.5) and non-bonded state (ELF < 0.4), the difference between the two states seem much less clear on the extreme than what was found in CO_2. Furthermore a middle region between 1.5Å and 2.5Å appears, although it is much less clear than the two others. Its ELF values are relatively high, or at least higher than those of the non-bonded states: 0.2<ELF<0.8 and it seems to indicate some sort of intermediary bonding state. It is very likely that this region signals hydrogen bonding, due to those characteristics. This analysis is the also true for the N-H bonds, where the same intermediary region appears.

For both C-N and C-O bonding, we mainly see the bonding states, but we do see that they are both also signaled with high value of the ELF (ELF > 0.5) although, we do see that the spread of ELF may not allow the ELF as the only criterion for bonding. As it

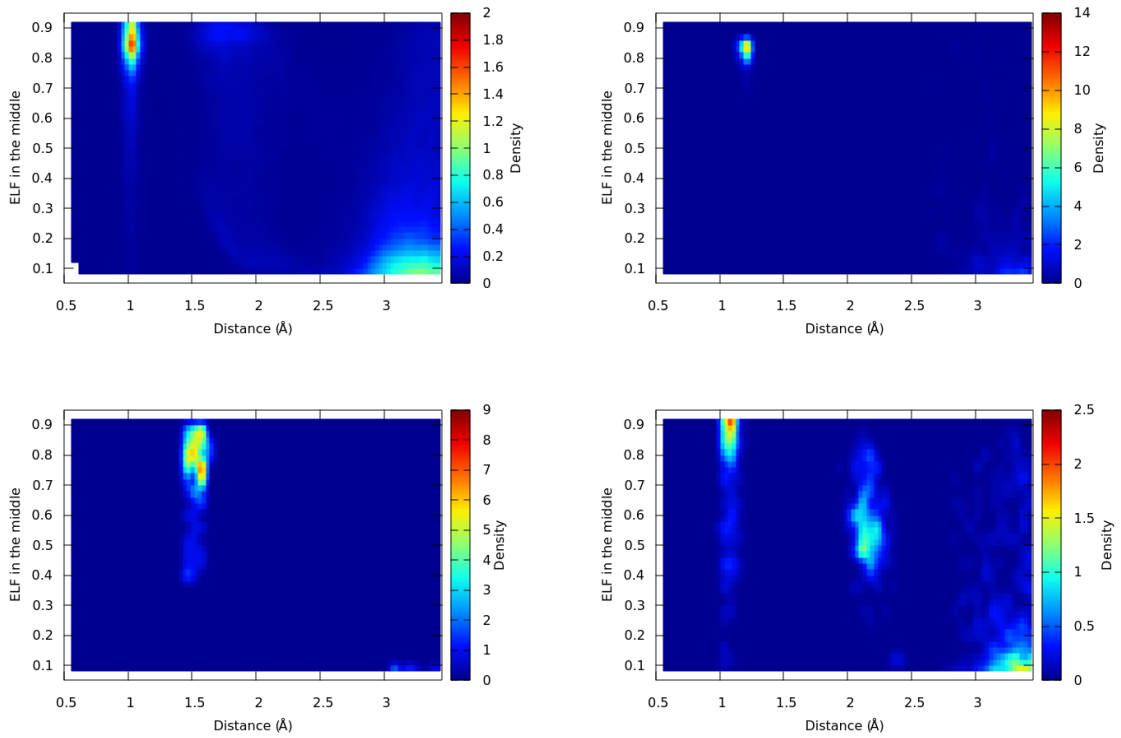


Figure B.1: Heatmap detailing the distribution of ELF value in the middle of the bond between various types of atom as a function of the distance between those atoms: O-H (top-left), C-O (top-right), C-N (bottom-left), N-H (bottom-right)

stands, and given our results for CO_2, there are two explination that are not mutually exclusive: the limited meshing of the ELF implies that we may miss the exact middle point, and in so-doing have a lower value than that of the actual middle point if it would have been sampled; the middle point value of the ELF between two atoms does may not contain enough the information about the bonding state, the maximum being shifted closer the either of the O or N atoms.

Appendix C

Test of the Permutation Invariant Vector (PIV)

During the course of this thesis we wondered if we could find a way to test the validity of a collective variable using the results from molecular dynamics simulation. One of the answer that we found was inspired by the analysis done in 8, to analyze the distribution of points in a plane with the differences in energy on the one hand and PIV distance on the other.

The rationale is the following: if the metric is valid, for small PIV distances (that is structures that should be very similar) there should not be very large difference in energy between the structures. We also would assume that this difference in energy should increase if not linearly at least smoothly with the PIV distance between structures.

In order to test the Permutation Invariant Vector with this method we computed the PIV distances between all pairs of structures (using piv_clustering [54]) in two different *ab initio* molecular dynamics simulations of liquid CO_2. Those simulation are a good test as they present two type of very different behavior: a molecular liquid on the one hand and a polymeric liquid on the other. Further, as PIV is reputed to be more efficient when long range environment is taken into account those systems being small (the simulation box is cubic with sides of less than 10Å long) also constitute very difficult case for the metric.

In C.1, the results show the results of the analysis for various parameters of the switching function:

$$\sigma(d_{i,j}) = \frac{1 - \left(\frac{d_{i,j}}{d_0}\right)^n}{1 - \left(\frac{d_{i,j}}{d_0}\right)^m} \tag{C.1}$$

We see that we see that the metric work particularly well in the molecular liquid case, where the maximum energy difference increases almost linearly with increasing PIV distance, at least in the case of moderate PIV distances. The results are still overall good in the case of the polymeric liquid, although the increase seem much more abrupt than

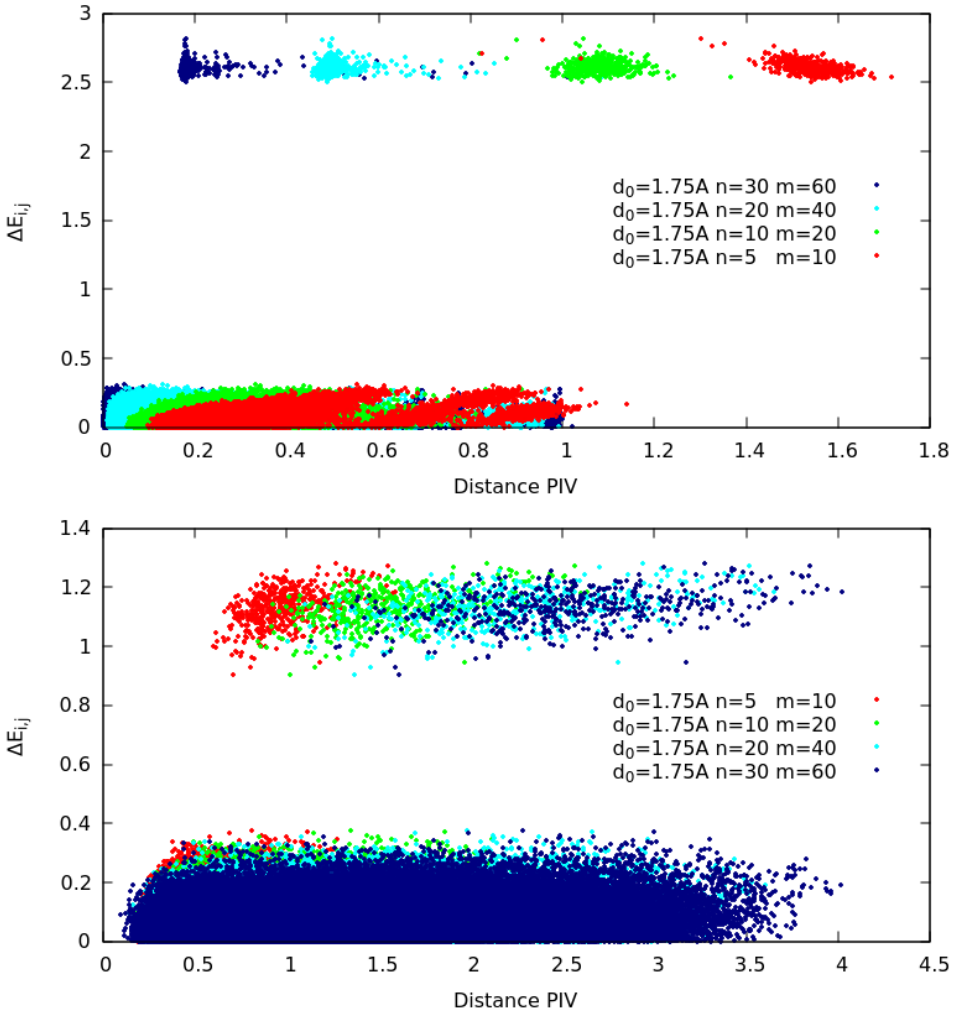


Figure C.1: PIV distance between all pairs of structure as a function of their difference in energy (in eV/CO_2). **Top** in the molecular case (3000K,$\sim$ 30GPa), **bottom** in the polymeric case (3000K,$\sim$ 70GPa)

in the molecular case. Another conclusion is that at least in the molecular case, a proper choice of the switching functions parameters is very important in order to obtain good results.

Appendix D

NH_3-H_2O Ima2 under high pressure

This project is a follow-up on Adrien Mafety's PhD thesis, which focused on the study of ammonia and water mixes under very high pressure (40-150GPa). Although we reintroduce some context, we redirect the reader to his thesis [53] or the one of Jean-Antoine Queyroux [186] for an experimental point of view of the same systems.

In this project we studied a specific phase made of a mixed molecular ice of NH_3-H_2O. This kind of system is of high interest in the experimental high pressure field as such mixed crystals may be found on planet-like bodies such as Titan [187] and Encelade [188]. Therefore understanding the structure and properties of those phases may prove invaluable in order to further our knowledge of those celestial bodies.

In this small chapter, we focus exclusively on a theoretically predicted phase Ima2. This phase, which is visible in figure D.1 was obtained by Adrien Mafety through the use of AIRSS (cf. 4.3.1), and it was predicted to be the most stable phase of the mix above 50GPa at 0K.

This phase exhibit an interesting structure where H_2O molecules are arranged so that they form a zig-zag line, with hydrogen almost shared between two successive oxygen. During the course of [53] it was found that the O-H-O distance would symmetrize starting at 80GPa. This symmetrization is of high interest as it may lead to exotic properties. The objective of this projet was therefore to study this symmetrization in order to find spectroscopic markers of this transition that could be observed experimentally.

In order to do so we relaxed the Ima2 structures at pressures ranging from 50 to 125GPa using the pw.x code in the Quantum Espresso [178] and then the infrared and Raman spectrum using the ph.x and dynmat.x codes of the same software. The calculations were made at the DFT level of theory using PBE functionnals and Vanderbilt[183] pseudopotentials from the Quantum Espresso pseudpotential library[178], the electronic convergence threshold was put at 10^{-12}, and we used a cut-off on the kinetic energy of 100 Ry for the wavefunction and 800 Ry for the electronic density. The results are shown in Figures D.2 and D.3.

In order to analyze the signature modes of the symmetrization, we tracked the resulting

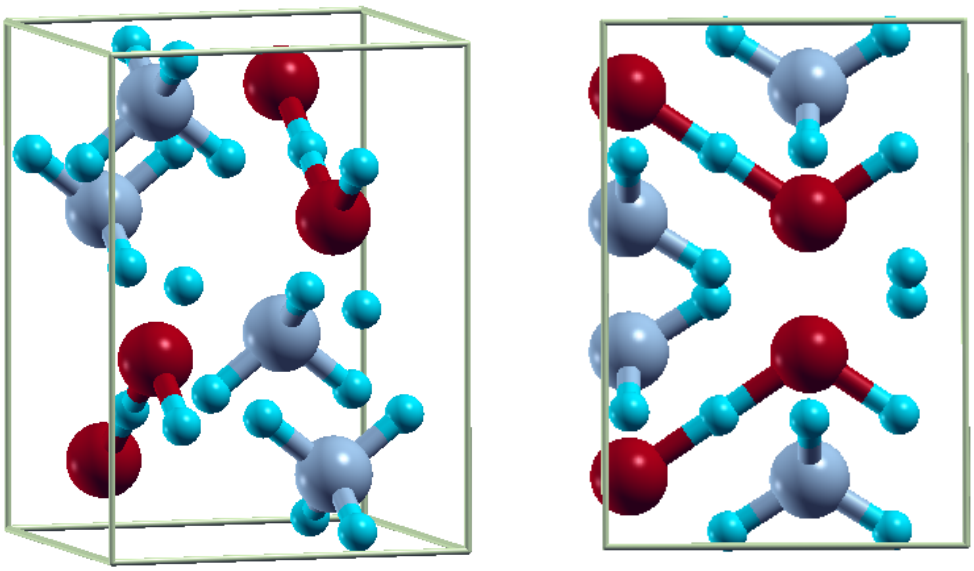


Figure D.1: View of the NH_3-H_2O Ima2 crystal phase. Oxygen are red, hydrogen are cyan, nitrogen are violet

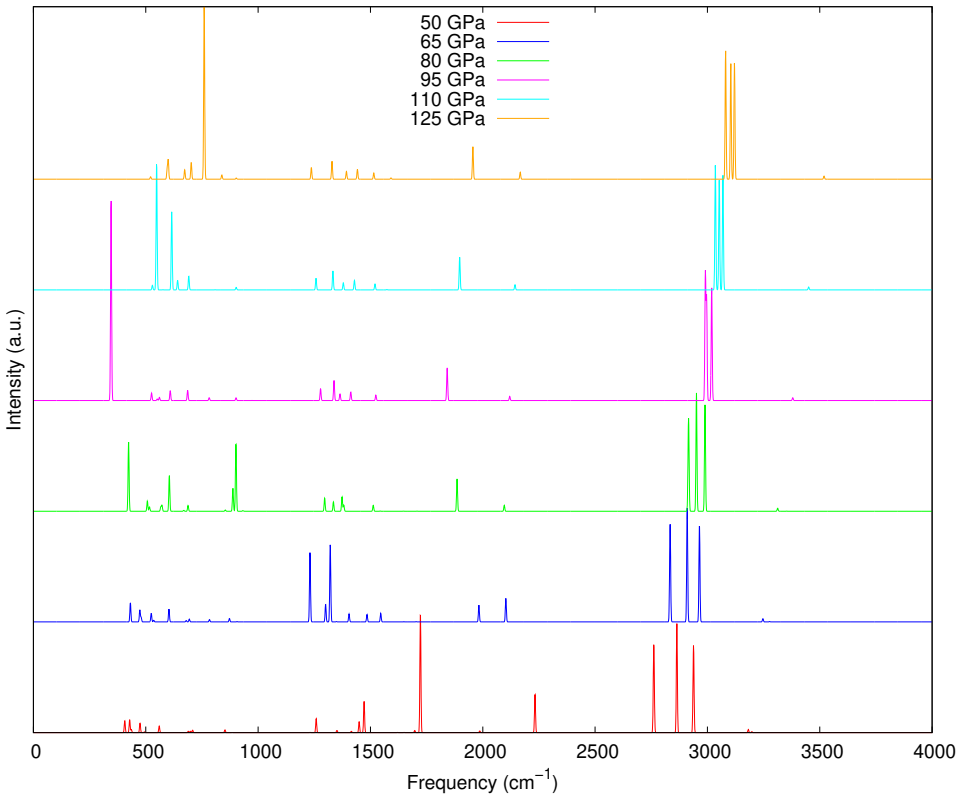


Figure D.2: Infrared spectrum of the NH_3H_2O Ima2 phase as a function of the pressure (GPa)

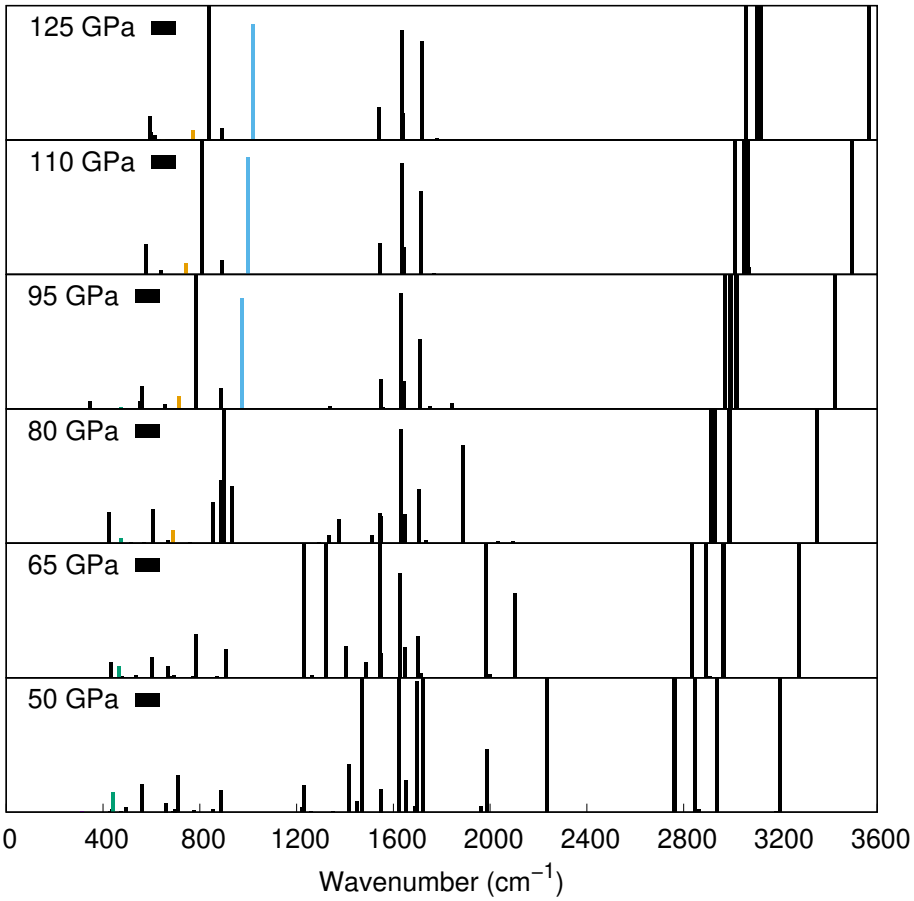


Figure D.3: Raman spectrum as a function of the pressure in GPa. The signature modes of the transition are colored for the non-symmetric phase (green) and symmetric phase (orange and red)

modes and found one mode that disapeared with the simmetryzation (green in D.3) and three other that appeared with it. Interestingly, all three modes that appear with the symmetrization share the common feature of vibrations on the O-H-O line: the oxygen atoms vibrate perpendicular to the H-O-H plane, alternating direction along the line as in figure D.4.

Another potential interesting aspects is the evoltion of the frequency of the modes of the structure with increasing pressure, as in figure D.5. Aside from the apparition and disapearing of the signature mode we do not find in the raman spectrum, any change of slope in the modes, which could be used as a signature of the transition.

The infrared spectrum does not give much information about the transition, contrary to the raman spectrum. The one exception being a change of slope in the dispersion relation of a stretching mode (see Figure D.6) at the transition pressure for a specific mode.

We therefore find specific vibrational signals that identify the symmetrization of the Ima2 phase, that can be experimentally measurable. Although theoretical calculations

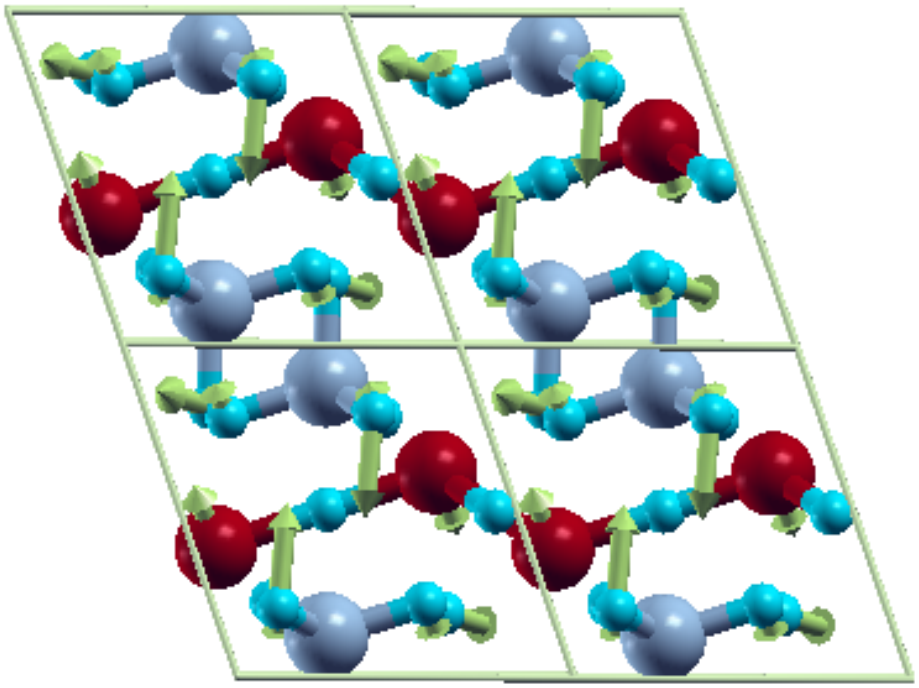


Figure D.4: Visualisation of the signature mode ($k \sim 600$ cm^{-1}) of the transition in the raman spectrum. Oxygen atoms are in red, nitrogen are violet and hydrogen cyan. The green arrow represent the forces.

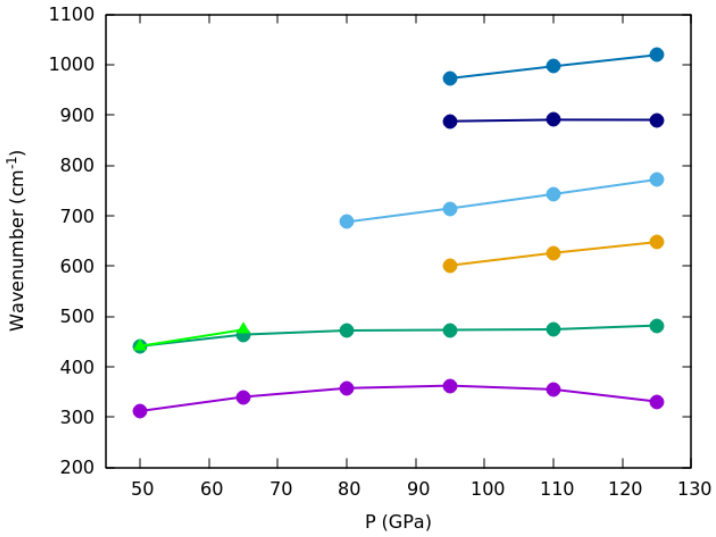


Figure D.5: Frequency of the Raman modes of Ima2 in the 0-1200 cm^{-1} region as a function of the pressure.

predict that the Ima2 phase is the most stable structure over 50GPa for NH_3-H_2O crystals, no experiments have currently been able to find any experimental evidence of its existence.

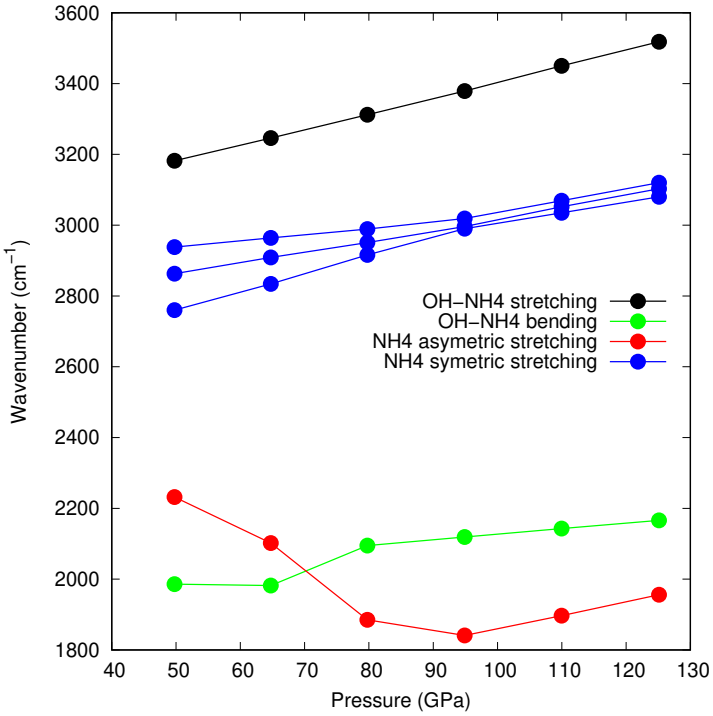


Figure D.6: Frequency of the IR modes of NH_3-H_2O-Ima2 as a function of the pressure in the 1800-3200 cm^{-1} range.

Bibliography

[1] G. Chiodini, F. Frondini, and F. Ponziani, "Deep structures and carbon dioxide degassing in central italy," *Geothermics*, vol. 24, no. 1, pp. 81–94, 1995.

[2] J. B. Lowenstern, "Carbon dioxide in magmas and implications for hydrothermal systems," *Mineralium Deposita*, vol. 36, no. 6, pp. 490–502, 2001.

[3] R. Dasgupta and M. M. Hirschmann, "The deep carbon cycle and melting in earth's interior," *Earth and Planetary Science Letters*, vol. 298, no. 1-2, pp. 1–13, 2010.

[4] F. Maeda, E. Ohtani, S. Kamada, T. Sakamaki, N. Hirao, and Y. Ohishi, "Diamond formation in the deep lower mantle: A high-pressure reaction of mgco 3 and sio 2," *Scientific reports*, vol. 7, p. 40602, 2017.

[5] X. Li, Z. Zhang, J.-F. Lin, H. Ni, V. B. Prakapenka, and Z. Mao, "New high-pressure phase of caco3 at the topmost lower mantle: Implication for the deep-mantle carbon transportation," *Geophysical Research Letters*, vol. 45, no. 3, pp. 1355–1360, 2018.

[6] F. Datchi, G. Weck, A. Saitta, Z. Raza, G. Garbarino, S. Ninet, D. Spaulding, J. Queyroux, and M. Mezouar, "Structure of liquid carbon dioxide at pressures up to 10 gpa," *Physical Review B*, vol. 94, no. 1, p. 014201, 2016.

[7] B. Boates, A. M. Teweldeberhan, and S. A. Bonev, "Stability of dense liquid carbon dioxide," *Proceedings of the National Academy of Sciences*, vol. 109, no. 37, pp. 14 808–14 812, 2012.

[8] K. D. Litasov, A. F. Goncharov, and R. J. Hemley, "Crossover from melting to dissociation of co2 under pressure: Implications for the lower mantle," *Earth and Planetary Science Letters*, vol. 309, no. 3-4, pp. 318–323, 2011.

[9] O. Tschauner, H.-k. Mao, and R. J. Hemley, "New transformations of co 2 at high pressures and temperatures," *Physical Review Letters*, vol. 87, no. 7, p. 075701, 2001.

[10] O. Maass and W. Barnes, "Some thermal constants of solid and liquid carbon dioxide," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 111, no. 757, pp. 224–244, 1926.

[11] W. Keesom and J. Köhler, "New determination of the lattice constant of carbon dioxide," *Physica*, vol. 1, no. 1-6, pp. 167–174, 1934.

[12] ——, "The lattice constant and expansion coefficient of solid carbon dioxide," *Physica*, vol. 1, no. 7-12, pp. 655–658, 1934.

[13] L.-g. Liu, "Dry ice ii, a new polymorph of co2," *Nature*, vol. 303, no. 5917, p. 508, 1983.

[14] B. Kuchta and R. Etters, "Prediction of a high-pressure phase transition and other properties of solid co 2 at low temperatures," *Physical Review B*, vol. 38, no. 9, p. 6265, 1988.

[15] H. Olijnyk, H. Däufer, H.-J. Jodl, and H. Hochheimer, "Effect of pressure and temperature on the raman spectra of solid co2," *The Journal of chemical physics*, vol. 88, no. 7, pp. 4204–4212, 1988.

[16] K. Aoki, H. Yamawaki, M. Sakashita, Y. Gotoh, and K. Takemura, "Crystal structure of the high-pressure phase of solid co2," *Science*, vol. 263, no. 5145, pp. 356–358, 1994.

[17] V. Iota and Y. C. S., "Phase diagram of carbon dioxide: Evidence for a new associated phase." *Physical review letters*, vol. 86, no. 26, p. 5922, 2001.

[18] F. Datchi, B. Mallick, A. Salamat, G. Rousse, S. Ninet, G. Garbarino, P. Bouvier, and M. Mezouar, "Structure and compressibility of the high-pressure molecular phase ii of carbon dioxide," *Physical Review B*, vol. 89, no. 14, p. 144101, 2014.

[19] R. Etters and B. Kuchta, "Static and dynamic properties of solid co2 at various temperatures and pressures," *The Journal of Chemical Physics*, vol. 90, no. 8, pp. 4537–4541, 1989.

[20] C.-S. Yoo, V. Iota, , and H. Cynn., "Nonlinear carbon dioxide at high pressures and temperatures," *Physical review letters*, vol. 86, no. 3, p. 444, 2001.

[21] J. H. Park, C. S. Yoo, V. Iota, H. Cynn, M. F. Nicol, and T. Le Bihan, "Crystal structure of bent carbon dioxide phase iv," *Physical Review B*, vol. 68, no. 1, p. 014107, 2003.

[22] F. A. Gorelli, V. M. Giordano, P. R. Salvi, and R. Bini, "Linear carbon dioxide in the high-pressure high-temperature crystalline phase iv," *Physical review letters*, vol. 93, no. 20, p. 205503, 2004.

[23] F. Datchi, V. M. Giordano, P. Munsch, and A. M. Saitta, "Structure of carbon dioxide phase iv: Breakdown of the intermediate bonding state scenario," *Physical review letters*, vol. 103, no. 18, p. 185701, 2009.

[24] V. Giordano and F. Datchi, "Molecular carbon dioxide at high pressure and high temperature," *EPL (Europhysics Letters)*, vol. 77, no. 4, p. 46002, 2007.

[25] V. Iota, C. Yoo, and H. Cynn, "Quartzlike carbon dioxide: an optically nonlinear extended solid at high pressures and temperatures," *Science*, vol. 283, no. 5407, pp. 1510–1513, 1999.

[26] C. Yoo, H. Cynn, F. Gygi, G. Galli, V. Iota, M. Nicol, S. Carlson, D. Häusermann, and C. Mailhiot, "Crystal structure of carbon dioxide at high pressure:"superhard" polymeric carbon dioxide," *Physical Review Letters*, vol. 83, no. 26, p. 5527, 1999.

[27] S. Serra, C. Cavazzoni, G. Chiarotti, S. Scandolo, and E. Tosatti, "Pressure-induced solid carbonates from molecular co2 by computer simulation," *Science*, vol. 284, no. 5415, pp. 788–790, 1999.

[28] B. Holm, R. Ahuja, A. Belonoshko, and B. Johansson, "Theoretical investigation of high pressure phases of carbon dioxide," *Physical Review Letters*, vol. 85, no. 6, p. 1258, 2000.

[29] J. Dong, J. K. Tomfohr, and O. F. Sankey, "Rigid intertetrahedron angular interaction of nonmolecular carbon dioxide solids," *Physical Review B*, vol. 61, no. 9, p. 5967, 2000.

[30] A. R. Oganov, S. Ono, Y. Ma, C. W. Glass, and A. Garcia, "Novel high-pressure structures of mgco3, caco3 and co2 and their role in earth's lower mantle," *Earth and Planetary Science Letters*, vol. 273, no. 1-2, pp. 38–47, 2008.

[31] M. Santoro, F. A. Gorelli, R. Bini, J. Haines, O. Cambon, C. Levelut, J. A. Montoya, and S. Scandolo, "Partially collapsed cristobalite structure in the non molecular phase v in co2," *Proceedings of the National Academy of Sciences*, vol. 109, no. 14, pp. 5176–5179, 2012.

[32] F. Datchi, B. Mallick, A. Salamat, and S. Ninet, "Structure of polymeric carbon dioxide co 2- v," *Physical Review Letters*, vol. 108, no. 12, p. 125701, 2012.

[33] V. Iota, C.-S. Yoo, J.-H. Klepeis, Z. Jenei, W. Evans, and H. Cynn, "Six-fold coordinated carbon dioxide vi," *Nature materials*, vol. 6, no. 1, p. 34, 2007.

[34] A. Togo, F. Oba, and I. Tanaka, "Transition pathway of c o 2 crystals under high pressures," *Physical Review B*, vol. 77, no. 18, p. 184101, 2008.

[35] J. Sun, D. D. Klug, R. Martoňák, J. A. Montoya, M.-S. Lee, S. Scandolo, and E. Tosatti, "High-pressure polymeric phases of carbon dioxide," *Proceedings of the National Academy of Sciences*, vol. 106, no. 15, pp. 6077–6081, 2009.

[36] M.-S. Lee, J. A. Montoya, and S. Scandolo, "Thermodynamic stability of layered structures in compressed co 2," *Physical Review B*, vol. 79, no. 14, p. 144102, 2009.

[37] A. Teweldeberhan, B. Boates, and S. Bonev, "Co2 in the mantle: Melting and solid–solid phase boundaries," *Earth and Planetary Science Letters*, vol. 373, pp. 228–232, 2013.

[38] M. Santoro, F. A. Gorelli, R. Bini, G. Ruocco, S. Scandolo, and W. A. Crichton, "Amorphous silica-like carbon dioxide," *Nature*, vol. 441, no. 7095, p. 857, 2006.

[39] M. Santoro and F. Gorelli, "High pressure solid state chemistry of carbon dioxide," *Chemical Society Reviews*, vol. 35, no. 10, pp. 918–931, 2006.

[40] J. A. Montoya, R. Rousseau, M. Santoro, F. Gorelli, and S. Scandolo, "Mixed threefold and fourfold carbon coordination in compressed co 2," *Physical review letters*, vol. 100, no. 16, p. 163002, 2008.

[41] J.-H. Park, C. Yoo, V. Iota, H. Cynn, M. Nicol, and T. Le Bihan, "Crystal structure of bent carbon dioxide phase iv," *Physical Review B*, vol. 68, no. 1, p. 014107, 2003.

[42] C. Yoo, H. Kohlmann, H. Cynn, M. Nicol, V. Iota, and T. LeBihan, "Crystal structure of pseudo-six-fold carbon dioxide phase ii at high pressures and temperatures," *Physical Review B*, vol. 65, no. 10, p. 104103, 2002.

[43] S. Bonev, F. Gygi, T. Ogitsu, and G. Galli, "High-pressure molecular phases of solid carbon dioxide," *Physical review letters*, vol. 91, no. 6, p. 065501, 2003.

[44] Y. Han, J. Liu, L. Huang, X. He, and J. Li, "Predicting the phase diagram of solid carbon dioxide at high pressure from first principles," *npj Quantum Materials*, vol. 4, no. 1, p. 10, 2019.

[45] V. M. Giordano, F. Datchi, and D. Agnès, "Melting curve and fluid equation of state of carbon dioxide at high pressure and high temperature." *The Journal of chemical physics*, vol. 125, no. 5, p. 054504, 2006.

[46] W. Sontising, Y. N. Heit, J. L. McKinley, and G. J. Beran, "Theoretical predictions suggest carbon dioxide phases iii and vii are identical," *Chemical science*, vol. 8, no. 11, pp. 7374–7382, 2017.

[47] J. Contreras-Garcia, A. M. Pendás, B. Silvi, and J. Recio, "Bases for understanding polymerization under pressure: the practical case of co2," *The Journal of Physical Chemistry B*, vol. 113, no. 4, pp. 1068–1073, 2009.

[48] X. Yong, H. Liu, M. Wu, Y. Yao, S. T. John, R. Dias, and C.-S. Yoo, "Crystal structures and dynamical properties of dense co2," *Proceedings of the National Academy of Sciences*, vol. 113, no. 40, pp. 11 110–11 115, 2016.

[49] K. F. Dziubek, M. Ende, D. Scelta, R. Bini, M. Mezouar, G. Garbarino, and R. Miletich, "Crystalline polymeric carbon dioxide stable at megabar pressures," *Nature communications*, vol. 9, no. 1, p. 3148, 2018.

[50] M. Santoro and F. A. Gorelli, "Constraints on the phase diagram of nonmolecular co 2 imposed by infrared spectroscopy," *Physical Review B*, vol. 80, no. 18, p. 184109, 2009.

[51] I. Gimondi and M. Salvalaglio, "Co2 packing polymorphism under pressure: Mechanism and thermodynamics of the i-iii polymorphic transition," *The Journal of chemical physics*, vol. 147, no. 11, p. 114502, 2017.

[52] F. Mouhat, "Fully quantum dynamics of protonated water clusters," Theses, Sorbonne Université, Sep. 2018. [En ligne]. Disponible: https://tel.archives-ouvertes.fr/tel-02137552

[53] A. Mafety, "Ab initio study of fluorinated and monohydrated ammonia ices under extreme thermodynamic conditions," Theses, Université Pierre et Marie Curie - Paris VI, Sep. 2016. [En ligne]. Disponible: https://tel.archives-ouvertes.fr/tel-01497646

[54] G. A. Gallet and F. Pietrucci, "Structural cluster analysis of chemical reactions in solution," *The Journal of chemical physics*, vol. 139, no. 7, p. 074101, 2013.

[55] F. Pietrucci and W. Andreoni, "Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale," *Physical review letters*, vol. 107, no. 8, p. 085504, 2011.

[56] A. Savin, O. Jepsen, J. Flad, O. K. Andersen, H. Preuss, and H. G. von Schnering, "Electron localization in solid-state structures of the elements – the diamond structure," *Angewandte Chemie*, vol. 31, no. 2, pp. 187–188, 1992.

[57] A. Rodriguez, M. d'Errico, E. Facco, and A. Laio, "Computing the free energy without collective variables," *Journal of chemical theory and computation*, vol. 14, no. 3, pp. 1206–1215, 2018.

[58] F. Tassone, G. L. Chiarotti, R. Rousseau, S. Scandolo, and E. Tosatti, "Dimerization of co2 at high pressure and temperature," *ChemPhysChem*, vol. 6, no. 9, pp. 1752–1756, 2005.

[59] D. Plašienka and R. Martoňák, "Structural evolution in high-pressure amorphous co 2 from ab initio molecular dynamics," *Physical Review B*, vol. 89, no. 13, p. 134105, 2014.

[60] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," *The Journal of chemical physics*, vol. 134, no. 17, p. 174105, 2011.

[61] C. J. Pickard and R. Needs, "Ab initio random structure searching," *Journal of Physics: Condensed Matter*, vol. 23, no. 5, p. 053201, 2011.

[62] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[63] G. N. Plass, "Effect of carbon dioxide variations on climate," *American journal of physics*, vol. 24, no. 5, pp. 376–387, 1956.

[64] G. S. Callendar, "Can carbon dioxide influence climate?" *Weather*, vol. 4, no. 10, pp. 310–314, 1949.

[65] A. Sengupta, M. Kim, C.-S. Yoo, and J. S. Tse, "Polymerization of carbon dioxide: A chemistry view of molecular-to-nonmolecular phase transitions," *The Journal of Physical Chemistry C*, vol. 116, no. 3, pp. 2061–2067, 2011.

[66] M. Santoro, J.-f. Lin, H.-k. Mao, and R. J. Hemley, "In situ high pt raman spectroscopy and laser heating of carbon dioxide," *The Journal of chemical physics*, vol. 121, no. 6, pp. 2780–2787, 2004.

[67] B. Boates, S. Hamel, E. Schwegler, and S. Bonev, "Structural and optical properties of liquid co2 for pressures up to 1 tpa," *The Journal of chemical physics*, vol. 134, no. 6, p. 064504, 2011.

[68] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln." 1927.

[69] F. Finocchi, "Density functionnal theory for beginners," 2011. [En ligne]. Disponible: https://tel.archives-ouvertes.fr/tel-01497646

[70] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Physical review*, vol. 140, no. 4A, p. A1133, 1965.

[71] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical review*, vol. 146, no. 3B, p. B864, 1964.

[72] J. Perdew, K. Burke, and M. Ernzerhof, "Perdew, burke, and ernzerhof reply," *Physical Review Letters*, vol. 80, no. 4, p. 891, 1998.

[73] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," *Physical review A*, vol. 38, no. 6, p. 3098, 1988.

[74] C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of the electron density," *Physical review B*, vol. 37, no. 2, p. 785, 1988.

[75] S. Grimme, "Accurate description of van der waals complexes by density functional theory including empirical corrections," *Journal of computational chemistry*, vol. 25, no. 12, pp. 1463–1473, 2004.

[76] S. Gohr, S. Grimme, T. Söhnel, B. Paulus, and P. Schwerdtfeger, "Pressure dependent stability and structure of carbon dioxide—a density functional study including long-range corrections," *The Journal of chemical physics*, vol. 139, no. 17, p. 174501, 2013.

[77] D. R. Hamann, M. Schlüter, and C. C., "Norm-conserving pseudopotentials," *Physical Review Letters*, vol. 43, no. 20, p. 1494, 1979.

[78] D. Vanderbilt, "Soft self-consistent pseudopotentials in a generalized eigenvalue formalism." *Physical review B*, vol. 41, no. 11, p. 7892, 1990.

[79] R. P. Feynman, "Forces in molecules," *Physical Review*, vol. 56, no. 4, p. 340, 1939.

[80] C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," *Math. Comput.*, vol. 19, no. 92, pp. 577–593, 1965.

[81] R. M. Martin and R. M. Martin, *Electronic structure: basic theory and practical methods*. Cambridge university press, 2004.

[82] A. D. Becke and E. K. E., "A simple measure of electron localization in atomic and molecular systems," *The Journal of chemical physics*, vol. 92, no. 9, pp. 5397–5403, 1990.

[83] P. Fuentealba, E. Chamorro, and S. J. C., "Understanding and using the electron localization function," *Theoretical and Computational Chemistry.*, vol. 19, pp. 57–85, 2007.

[84] C. V. Weizsäcker, "Zur theorie der kernmassen," *Zeitschrift für Physik A Hadrons and Nuclei*, vol. 96, no. 7, pp. 431–458, 1935.

[85] J. Contreras-Garcia and J. M. Recio, "Electron delocalization and bond formation under the elf framework," *Theoretical Chemistry Accounts*, vol. 128, no. 4-6, pp. 411–418, 2011.

[86] M. E. Alikhani, F. Fuster, and S. B., "What can tell the topological analysis of elf on hydrogen bonding?" *Structural Chemistry*, vol. 16, no. 3, pp. 203–210, 2005.

[87] B. Silvi and R. Henryk, "Hydrogen bonding and delocalization in the elf analysis approach." *Physical Chemistry Chemical Physics*, vol. 18, no. 39, pp. 27 442–27 449, 2016.

[88] F. Fuster and S. Bernard, "Does the topological approach characterize the hydrogen bond?" *Nature*, vol. 371, no. 6499, 1994.

[89] B. Silvi and S. Andreas, "Classification of chemical bonds based on topological analysis of electron localization functions," *Theoretical Chemistry Accounts*, vol. 104, no. 1, pp. 13–21, 2000.

[90] A. Savin, R. Nesper, S. Wengert, and T. F. Fässler, "Elf: The electron localization function," *Angewandte Chemie*, vol. 36, no. 17, pp. 1808–1832, 1997.

[91] J. Lennard, "On the determination of molecular fields," *Proc. R. Soc. London*, vol. 106, pp. 441–477, 1924.

[92] Z. Zhang and Z. Duan, "An optimized molecular potential for carbon dioxide," *The Journal of chemical physics*, vol. 122, no. 21, p. 214507, 2005.

[93] X. Huang, B. J. Braams, and J. M. Bowman, "Ab initio potential energy and dipole moment surfaces for h 5 o 2+," *The Journal of chemical physics*, vol. 122, no. 4, p. 044308, 2005.

[94] A. B. McCoy, X. Huang, S. Carter, M. Y. Landeweer, and J. M. Bowman, "Full-dimensional vibrational calculations for h 5 o 2+ using an ab initio potential energy surface," 2005.

[95] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Physical review letters*, vol. 98, no. 14, p. 146401, 2007.

[96] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *The Journal of chemical physics*, vol. 134, no. 7, p. 074106, 2011.

[97] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, "Bypassing the kohn-sham equations with machine learning," *Nature communications*, vol. 8, no. 1, p. 872, 2017.

[98] L. Verlet, "Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules," *Physical review*, vol. 159, no. 1, p. 98, 1967.

[99] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, "Lincs: a linear constraint solver for molecular simulations," *Journal of computational chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[100] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "Gromacs: fast, flexible, and free," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[101] M. Tuckerman, *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010.

[102] H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola, and J. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of chemical physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[103] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *The Journal of chemical physics*, vol. 126, no. 1, p. 014101, 2007.

[104] S. Nosé, "A molecular dynamics method for simulations in the canonical ensemble," *Molecular physics*, vol. 52, no. 2, pp. 255–268, 1984.

[105] ——, "A unified formulation of the constant temperature molecular dynamics methods," *The Journal of chemical physics*, vol. 81, no. 1, pp. 511–519, 1984.

[106] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Physical review A*, vol. 31, no. 3, p. 1695, 1985.

[107] M. Parrinello and A. Rahman, "Crystal structure and pair potentials: A molecular-dynamics study," *Physical Review Letters*, vol. 45, no. 14, p. 1196, 1980.

[108] R. Martoňák, A. Laio, and M. Parrinello, "Predicting crystal structures: the parrinello-rahman method revisited," *Physical review letters*, vol. 90, no. 7, p. 075503, 2003.

[109] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058301, 2012.

[110] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Physical Review B*, vol. 87, no. 18, p. 184115, 2013.

[111] B. Bollobás, *Modern Graph Theory*. Springer, 1998.

[112] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello, "Plumed: A portable plugin for free-energy calculations with molecular dynamics," *Comput. Phys. Commun*, vol. 180, no. 10, pp. 1961–1972, 2009.

[113] S. Pipolo, M. Salanne, G. Ferlat, S. Klotz, A. M. Saitta, and F. Pietrucci, "Navigating at will on the water phase diagram," *Physical review letters*, vol. 119, no. 24, p. 245701, 2017.

[114] S. Laporte, "The Electric Field at an Oxide Surface - Impact on Reactivity of Prebiotic Molecules," Theses, Université Pierre et Marie Curie - Paris VI, Sep. 2016. [En ligne]. Disponible: https://tel.archives-ouvertes.fr/tel-01469587

[115] A. Pérez-Villa and F. Pietrucci, "Free energy, friction, and mass profiles from short molecular dynamics trajectories," *arXiv preprint arXiv:1810.00713*, 2018.

[116] A. Pérez-Villa, A. M. Saitta, T. Georgelin, J.-F. Lambert, F. Guyot, M.-C. Maurel, and F. Pietrucci, "Synthesis of rna nucleotides in plausible prebiotic conditions from ab initio computer simulations," *The journal of physical chemistry letters*, vol. 9, no. 17, pp. 4981–4987, 2018.

[117] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[118] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Springer, 1971, pp. 134–151.

[119] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[120] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[121] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[122] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[123] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. Van Gunsteren, and A. E. Mark, "Peptide folding: when simulation meets experiment," *Angewandte Chemie International Edition*, vol. 38, no. 1-2, pp. 236–240, 1999.

[124] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[125] C. Rossant, S. N. Kadir, D. F. Goodman, J. Schulman, M. L. Hunter, A. B. Saleem, A. Grosmark, M. Belluscio, G. H. Denfield, A. S. Ecker *et al.*, "Spike sorting for large, dense electrode arrays," *Nature neuroscience*, vol. 19, no. 4, p. 634, 2016.

[126] J. Cao, J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers *et al.*, "Comprehensive single-cell transcriptional profiling of a multicellular organism," *Science*, vol. 357, no. 6352, pp. 661–667, 2017.

[127] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," *Physical Chemistry Chemical Physics*, vol. 18, no. 20, pp. 13 754–13 769, 2016.

[128] L. Yaohui, M. Zhengming, and Y. Fang, "Adaptive density peak clustering based on k-nearest neighbors with aggregating strategy," *Knowledge-Based Systems*, vol. 133, pp. 208–220, 2017.

[129] M. d'Errico, E. Facco, A. Laio, and A. Rodriguez, "Automatic topography of high-dimensional data sets by non-parametric density peak clustering," *arXiv preprint arXiv:1802.10549*, 2018.

[130] G. R. Bowman, V. S. Pande, and F. Noé, *An introduction to Markov state models and their application to long timescale molecular simulation.* Springer Science & Business Media, 2013, vol. 797.

[131] P. G. Doyle, C. M. Grinstead, and J. L. Snell, "Grinstead and snell's introduction to probability," 2006.

[132] F. Noé and S. Fischer, "Transition networks for modeling the kinetics of conformational change in macromolecules," *Current opinion in structural biology*, vol. 18, no. 2, pp. 154–162, 2008.

[133] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 562–12 566, 2002.

[134] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, and M. Parrinello, "Assessing the accuracy of metadynamics," *The journal of physical chemistry B*, vol. 109, no. 14, pp. 6714–6721, 2005.

[135] G. Bussi, A. Laio, and M. Parrinello, "Equilibrium free energies from nonequilibrium metadynamics," *Physical review letters*, vol. 96, no. 9, p. 090601, 2006.

[136] Y. Crespo, F. Marinelli, F. Pietrucci, and A. Laio, "Metadynamics convergence law in a multidimensional system," *Physical Review E*, vol. 81, no. 5, p. 055701, 2010.

[137] J. F. Dama, M. Parrinello, and G. A. Voth, "Well-tempered metadynamics converges asymptotically," *Physical review letters*, vol. 112, no. 24, p. 240602, 2014.

[138] P. Tiwary and M. Parrinello, "A time-independent free energy estimator for metadynamics," *The Journal of Physical Chemistry B*, vol. 119, no. 3, pp. 736–742, 2014.

[139] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: a smoothly converging and tunable free-energy method," *Physical review letters*, vol. 100, no. 2, p. 020603, 2008.

[140] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977.

[141] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.

[142] B. Roux, "The calculation of the potential of mean force using computer simulations," *Computer physics communications*, vol. 91, no. 1-3, pp. 275–282, 1995.

[143] F. Zhu and G. Hummer, "Convergence and error estimation in free energy calculations using the weighted histogram analysis method," *Journal of computational chemistry*, vol. 33, no. 4, pp. 453–465, 2012.

[144] E. Rosta and G. Hummer, "Free energies from dynamic weighted histogram analysis using unbiased markov state model," *Journal of chemical theory and computation*, vol. 11, no. 1, pp. 276–285, 2014.

[145] E. H. Thiede, B. Van Koten, J. Weare, and A. R. Dinner, "Eigenvector method for umbrella sampling enables error analysis," *The Journal of chemical physics*, vol. 145, no. 8, p. 084115, 2016.

[146] A. R. Oganov and C. W. Glass, "Crystal structure prediction using ab initio evolutionary techniques: Principles and applications," *J. Chem. Phys.*, vol. 124, no. 24, p. 244704, 2006.

[147] Y. Wang, J. Lv, L. Zhu, and Y. Ma, "Crystal structure prediction via particle-swarm optimization," *Physical Review B*, vol. 82, no. 9, p. 094116, 2010.

[148] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," *The Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[149] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, "Adaptive biasing force method for scalar and vector free energy calculations," *The Journal of chemical physics*, vol. 128, no. 14, p. 144120, 2008.

[150] J. Schlitter, M. Engels, and P. Krüger, "Targeted molecular dynamics: a new approach for searching pathways of conformational transitions," *Journal of molecular graphics*, vol. 12, no. 2, pp. 84–89, 1994.

[151] H. T. Stokes and D. M. Hatch, "Findsym: program for identifying the space-group symmetry of a crystal," *Journal of Applied Crystallography*, vol. 38, no. 1, pp. 237–238, 2005.

[152] M. Martinez-Canales, C. J. Pickard, and R. J. Needs, "Thermodynamically stable phases of carbon at multiterapascal pressures," *Physical review letters*, vol. 108, no. 4, p. 045704, 2012.

[153] C. J. Pickard, M. Martinez-Canales, and R. J. Needs, "Decomposition and terapascal phases of water ice," *Physical review letters*, vol. 110, no. 24, p. 245701, 2013.

[154] ——, "Density functional theory study of phase iv of solid hydrogen," p. 214114, 2012.

[155] N. Takafuji, K. Fujino, T. Nagai, Y. Seto, and D. Hamane, "Decarbonation reaction of magnesite in subducting slabs at the lower mantle," *Physics and Chemistry of Minerals*, vol. 33, no. 10, pp. 651–654, 2006.

[156] N. Troullier and J. L. Martins, "Efficient pseudopotentials for plane-wave calculations," *Physical review B*, vol. 43, no. 3, p. 1993, 1991.

[157] CPMD, Copyright IBM Corp 1990-2008, Copyright MPI für Festkörperforschung Stuttgart 1997-2001, "Cpmd." [En ligne]. Disponible: http://cpmd.org/

[158] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. Probert, K. Refson, and M. C. Payne, "First principles methods using castep," *Zeitschrift für Kristallographie-Crystalline Materials*, vol. 220, no. 5/6, pp. 567–570, 2005.

[159] C.-S. Yoo, "Physical and chemical transformations of highly compressed carbon dioxide at bond energies," *Physical Chemistry Chemical Physics*, vol. 15, no. 21, pp. 7949–7966, 2013.

[160] J. V. Lauritsen, J. Kibsgaard, S. Helveg, H. Topsøe, B. S. Clausen, E. Lægsgaard, and F. Besenbacher, "Size-dependent structure of mos 2 nanocrystals," *Nat. Nanotechnol.*, vol. 2, no. 1, p. 53, 2007.

[161] S. Helveg, J. V. Lauritsen, E. Lægsgaard, I. Stensgaard, J. K. Nørskov, B. Clausen, H. Topsøe, and F. Besenbacher, "Atomic-scale structure of single-layer mos 2 nanoclusters," *Phys. Rev. Lett.*, vol. 84, no. 5, p. 951, 2000.

[162] N. Bertram, Y. D. Kim, G. Ganteför, Q. Sun, P. Jena, J. Tamuliene, and G. Seifert, "Experimental and theoretical studies on inorganic magic clusters: M4x6 (m= w, mo, x= o, s)," *Chem. Phys. Lett.*, vol. 396, no. 4-6, pp. 341–345, 2004.

[163] G. Seifert, J. Tamuliene, and S. Gemming, "Mons2n+ x clusters—magic numbers and platelets," *Comput. Mater. Sci.*, vol. 35, no. 3, pp. 316–320, 2006.

[164] S. Gemming, G. Seifert, M. Götz, T. Fischer, and G. Ganteför, "Transition metal sulfide clusters below the cluster–platelet transition: Theory and experiment," *Phys. Status Solidi B*, vol. 247, no. 5, pp. 1069–1076, 2010.

[165] S. Gemming, J. Tamuliene, G. Seifert, N. Bertram, Y. D. Kim, and G. Ganteför, "Electronic and geometric structures of mo x s y and w x s y (x= 1, 2, 4; y= 1–12) clusters," *Appl. Phys. A*, vol. 82, no. 1, pp. 161–166, 2006.

[166] P. Murugan, V. Kumar, Y. Kawazoe, and N. Ota, "Ab initio study of structural stability of mo- s clusters and size specific stoichiometries of magic clusters," *J. Phys. Chem. A*, vol. 111, no. 14, pp. 2778–2782, 2007.

[167] ——, "Assembling nanowires from mo- s clusters and effects of iodine doping on electronic structure," *Nano Lett.*, vol. 7, no. 8, pp. 2214–2219, 2007.

[168] ——, "Atomic structures and magnetism in small mos 2 and ws 2 clusters," *Phys. Rev. A*, vol. 71, no. 6, p. 063203, 2005.

[169] B. Wang, N. Wu, X.-B. Zhang, X. Huang, Y.-F. Zhang, W.-K. Chen, and K.-N. Ding, "Probing the smallest molecular model of mos2 catalyst: S2 units in the mos n–/0 (n= 1–5) clusters," *J. Phys. Chem. A*, vol. 117, no. 27, pp. 5632–5641, 2013.

[170] D. D. J. Singh, T. Pradeep, J. Bhattacharjee, and U. Waghmare, "Novel cage clusters of mos2 in the gas phase," *J. Phys. Chem. A*, vol. 109, no. 33, pp. 7339–7342, 2005.

[171] N. J. Mayhall, E. L. Becher, III, A. Chowdhury, and K. Raghavachari, "Molybdenum oxides versus molybdenum sulfides: Geometric and electronic structures of mo3x y-(x= o, s and y= 6, 9) clusters," *J. Phys. Chem. A*, vol. 115, no. 11, pp. 2291–2296, 2011.

[172] I. Laraib, J. Karthikeyan, and P. Murugan, "First principles modeling of mo 6 s 9 nanowires via condensation of mo 4 s 6 clusters and the effect of iodine doping on structural and electronic properties," *Phys. Chem. Chem. Phys.*, vol. 18, no. 7, pp. 5471–5476, 2016.

[173] M. J. Patterson, J. M. Lightstone, and M. G. White, "Structure of molybdenum and tungsten sulfide m x s y+ clusters: Experiment and dft calculations," *J. Phys. Chem. A*, vol. 112, no. 47, pp. 12 011–12 021, 2008.

[174] R. Hafizi, S. J. Hashemifar, M. Alaei, M. Jangrouei, and H. Akbarzadeh, "Stable isomers and electronic, vibrational, and optical properties of ws2 nano-clusters: A first-principles study," *J. Chem. Phys.*, vol. 145, no. 21, p. 214303, 2016.

[175] Y.-Y. Wang, J.-J. Deng, X. Wang, J.-T. Che, and X.-L. Ding, "Small stoichiometric (mos 2) n clusters with the 1t phase," *Phys. Chem. Chem. Phys.*, vol. 20, no. 9, pp. 6365–6373, 2018.

[176] N. Bertram, J. Cordes, Y. D. Kim, G. Ganteför, S. Gemming, and G. Seifert, "Nanoplatelets made from mos2 and ws2," *Chem. Phys. Lett.*, vol. 418, no. 1-3, pp. 36–39, 2006.

[177] G. Rossi and R. Ferrando, "Searching for low-energy structures of nanoparticles: a comparison of different methods and algorithms," *J. Phys. Condens. Matter*, vol. 21, no. 8, p. 084208, 2009.

[178] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo *et al.*, "Quantum espresso: a modular and open-source software project for quantum simulations of materials," *Journal of physics: Condensed matter*, vol. 21, no. 39, p. 395502, 2009.

[179] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, no. 18, p. 3865, 1996.

[180] S. Goedecker, M. Teter, and J. Hutter, "Separable dual-space gaussian pseudopotentials," *Phys. Rev. B*, vol. 54, no. 3, p. 1703, 1996.

[181] M. Krack, "Pseudopotentials for h to kr optimized for gradient-corrected exchange-correlation functionals," *Theor. Chem. Acc.*, vol. 114, no. 1-3, pp. 145–152, 2005.

[182] A. M. Rappe, K. M. Rabe, E. Kaxiras, and J. Joannopoulos, "Optimized pseudopotentials," *Phys. Rev. B*, vol. 41, no. 2, p. 1227, 1990.

[183] D. Vanderbilt, "Optimally smooth norm-conserving pseudopotentials," *Physical Review B*, vol. 32, no. 12, p. 8412, 1985.

[184] F. Datchi, M. Moog, F. Pietrucci, and A. M. Saitta, "Polymeric phase v of carbon dioxide has not been recovered at ambient pressure and has a unique structure," *Proceedings of the National Academy of Sciences*, vol. 114, no. 5, pp. E656–E657, 2017.

[185] H. Liu, X. Yong, Y. Yao, S. T. John, and C.-S. Yoo, "Reply to datchi et al.: Recovered phase co2-v at low temperature and a newly predicted 3d-extended co2 phase," *Proceedings of the National Academy of Sciences*, vol. 114, no. 5, pp. E658–E659, 2017.

[186] J.-A. Queyroux, "Fusion, structure et diagramme de phases des glaces d'eau et d'ammoniac sous conditions extrêmes de pression et de température," Ph.D. dissertation, Paris 6, 2017.

[187] O. Grasset, C. Sotin, and F. Deschamps, "On the internal structure and dynamics of titan," *Planetary and Space Science*, vol. 48, no. 7-8, pp. 617–636, 2000.

[188] C. Porco, P. Helfenstein, P. Thomas, A. Ingersoll, J. Wisdom, R. West, G. Neukum, T. Denk, R. Wagner, T. Roatsch *et al.*, "Cassini observes the active south pole of enceladus," *science*, vol. 311, no. 5766, pp. 1393–1401, 2006.