Study of phase transformation of matter through topological coordinates

Alexandre Jedrecy

École Doctorale 397 : Physique et Chimie des Matériaux

Préparée à l'Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC)

Study of phase transformation of matter through topological coordinates

Par Alexandre Jedrecy

Thèse de doctorat de physique

Dirigée par Fabio Pietrucci & Marco Saitta

Présentée et soutenue publiquement le 14/10/2020.

Devant un jury composé de :

Présidente du jury :	Maria Barbi	LPTMC - Sorbonne Université
Rapporteurs :	Carlos Vega de las Heras	Universidad Complutense de Madrid
	Frédéric van Wijland	LMSC - Université de Paris
Directeurs :	Fabio Pietrucci	IMPMC - Sorbonne Université
	Marco Saitta	IMPMC - Sorbonne Université
Examinateurs :	Frédéric Caupin	ILM - Université Claude Bernard Lyon 1
	Michele Ceriotti	COSMO - EPFL

Table des matières

1 Résumé en français

Nous avons décidé d'écrire cette thèse en anglais, car tous les membres du jury ne sont pas francophones, mais aussi par soucis d'accessibilité au plus grand nombre. Ce choix est rendu possible par le règlement de l'école doctorale 397, sous réserve qu'un "*résumé substantiel* $\sim$ 10 *pages*" soit écrit en français. C'est ce résumé qui est présenté dans les pages suivantes, en suivant la structure du texte principal en anglais.

We choosed to write this thesis in english as some of the jury's members are not francophone, and to make it accessible to a greater number. This choice is possible in accordance to the doctoral school 397, if a substantial french summary of $\sim$ 10 pages is furnished. This summary is presented in the following, keeping the same structure as the main text in english.

1.1 Introduction

Si nous regardions les transformations de la matière condensée à une échelle atomique, une grande variété de mécanismes se déploierait sous nos yeux. Étudier les processus sous-jacents à ces transformations est un des plus vieux problèmes de la communauté qui étudie la matière condensée. À cause de la nature évanescente des états de transitions, les observations expérimentales sont difficiles à mener et souvent limitées dans leurs résultats. C'est pourquoi les études numériques ont joué un rôle majeur dans notre compréhension des processus de transformations dans la matière.

Curieusement, l'eau est à la fois un des matériaux le plus omniprésent et le plus bizarre sur Terre. Ses nombreuses anomalies comparées à d'autres liquides, notamment celles aux basses températures, font parties des problèmes les plus mal compris et les plus débattus dans les dernières décennies [1]. Même le problème fondamental de la nucléation de la glace, c'est-à-dire comment l'eau liquide se transforme en glace, n'est toujours pas bien compris, avec de nombreux travaux qui l'étudient encore [2].

D'un point de vue général, un système peut accéder à différents états physiques, pour un jeu de température et de pression donné. Par exemple pour des conditions ambiantes de température et de pression, l'eau peut se trouver sous des formes liquides, gazeuses, ou solides. Pour décrire la stabilité de ces états, historiquement, les physiciens ont introduit la notion d'énergie libre. Un état dont l'énergie libre est minimale est dit stable, ou métastable si l'énergie est minimale localement. Chaque minimum est séparé par une barrière d'énergie libre, qui empêche le système de transiter librement entre les différents états. Chaque minimum de l'énergie libre piège le système pour un temps long (de quelques microsecondes à quelques millions d'années), en comparaison du temps typique de vibration moléculaire (quelques femtosecondes). Rarement, le système va sauter par-dessus la barrière et transiter d'un état à un autre.

Pour simuler la matière, dans cette thèse nous nous sommes limités à une approche classique de dynamique moléculaire. Dans cette approche, toutes les molécules sont décrites par des modèles de champs de forces, pour lesquels les équations du mouvement sont résolues numériquement afin de calculer la vitesse et la position de chaque molécule à chaque pas de temps. En pratique, les simulations de dynamique

moléculaire classiques sont limitées par la puissance de calcul dont nous disposons, qui souvent se limite à l'ordre de quelques centaines de nanosecondes. Soit au minimum 6 ordres de grandeurs en dessous de l'échelle de temps typique d'une transition dans la matière condensée.

Pour pallier ce problème, une approche consiste à coupler la dynamique moléculaire avec des méthodes d'échantillonnage amélioré, qui permettent de réduire le temps nécessaire pour échantillonner une transition. Cette approche nécessite de définir des coordonnées de réactions, qui sont censées décrire une transition donnée. En pratique ce n'est pas trivial du tout de trouver des coordonnées de réactions optimales, qui souvent nécessitent des connaissances très précises du système et de la transition étudiée. Il y a un réel besoin de développer des méthodes générales et fiables pour automatiser ce processus, en réduisant les connaissances nécessaires au préalable pour étudier une transformation.

Nous avons donc deux problèmes : d'un côté le développement de méthodes générales pour étudier les transformations de la matière, et de l'autre l'étude de deux problèmes compliqués de l'eau. L'objectif de cette thèse est de montrer qu'il est possible de développer des métriques générales pour étudier les transformations, en les appliquant sur le très controversé problème de la transition liquide-liquide et sur le problème complexe de la nucléation homogène de la glace pour l'eau.

1.2 Simuler la matière condensée

Comme dit précédemment, pour simuler la matière condensée nous allons utiliser une approche classique, basée sur des modèles de champs de forces. Pour l'eau, il existe plusieurs dizaines de modèles différents, avec leurs forces et leurs faiblesses. Ici, nous allons parler plus spécifiquement de quatre modèles.

Le premier est le modèle mW [3]. Basé sur l'observation que l'eau et le silicium possèdent des propriétés structurelles similaires de par leur tendance à former des structures tétraédriques, il est une modification d'un modèle utilisé pour décrire le silicium. Concrètement, il modélise une molécule d'eau comme une boule avec un certain rayon, et dont l'arrangement en structure tétraédrique est créé artificiellement par l'ajout d'un terme quadratique qui favorise la formation d'un angle $\theta_0 = 109.47°$ entre triplets de molécules. Comme il ne contient que des termes à courte portée, il offre de très bonnes performances de calcul et décrit très correctement le diagramme de phase de l'eau et ses propriétés à des pressions atmosphériques, sauf la cinétique. Mais comme vous pouvez vous en douter, dès qu'on sort des gammes de pression atmosphérique, où la structure de l'eau n'est plus forcément tétraédrique, ses prédictions diffèrent largement des observations expérimentales en trouvant une phase qui n'existe pas.

Le deuxième est le modèle ST2 [4]. Ici l'eau est décrite comme une molécule H_2O, mais pour laquelle la charge négative serait localisée à la position de deux particules fictives, arrangées de manière à former un tétraèdre avec les atomes d'hydrogène, l'atome d'oxygène ne possédant pas de charge. Cette façon de décrire la molécule d'eau mène à sur-structurer l'eau, et donne de mauvaises prédictions pour son diagramme de phase et ses propriétés en dehors des basses pressions atmosphériques.

Les deux derniers modèles font partie d'une même famille, mais sont paramétrés différemment. Ils se basent sur la géométrie TIP4P, où la molécule d'eau est décrite

comme une molécule H_2O, mais pour laquelle la charge négative serait localisée à la position d'une particule fictive entre les deux atomes d'hydrogène, dans le plan formé par le trièdre H-O-H. Ce qui est différent est la façon dont ils ont été paramétrés. Le modèle TIP4P/2005 a été paramétré pour reproduire les données expérimentales de la courbe de densité maximale de l'eau [5]. Le modèle TIP4P/Ice a été paramétré pour reproduire la température de solidification de l'eau liquide à $T = 0°$ C, et les différentes densités des phases cristallines [6]. Ces deux modèles décrivent très correctement le diagramme de phase de l'eau, en trouvant correctement toutes ses phases, mais ils sont légèrement décalés par rapport aux données expérimentales. De plus leur cinétique est beaucoup plus réaliste que celle décrite par le modèle mW.

1.3 Décrire un système avec des variables collectives

Pour étudier les transformations de la matière, il nous faut des outils pour pouvoir distinguer les différents états qu'elle peut atteindre. L'eau par exemple a plus de 10 phases solides : comment pourrait-on les distinguer ? Avec la dynamique moléculaire, nous avons accès aux positions de chaque particule, mais ça représente vite un volume de données beaucoup trop important. L'idée des variables collectives est de calculer des données spécifiques pour un système particulier à partir de la position de ses particules. Par exemple, cette variable peut être le nombre de voisins d'une particule.

Quand une variable collective $s(x), x \in \mathbb{R}^{3N}$ est utilisée pour décrire une transition entre deux états ou plus, on parle de *paramètre d'ordre* et de *coordonnée de réaction*. En général le premier terme fait référence à une variable capable de distinguer les états localement stables, tandis que le second fait référence à la meilleure variable collective possible, qui décrit non seulement les états localement stables, mais aussi les détails du mécanisme de transition.

Concrètement, utiliser une variable collective $s(x)$ revient à projeter l'espace des configurations sur cette variable, en réduisant sa dimensionnalité. Une opération qui n'est pas triviale, mais qui rend les profils d'énergie libre plus simple à analyser. De plus, souvent cette projection donne un nouveau point de vue sur la transformation étudiée, avec son lot d'information sur la physique sous-jacente.

Deux grands types de variable collective existent : celles qui sont locales et nous renseignent sur l'état de chaque particule ; et celles qui sont globales et nous renseignent sur l'état du système dans son entier.

Parmi la multitude de variables collectives utilisées, une classe de variables basées sur un vecteur particulier a servi de fondation pour tout le travail effectué pendant cette thèse : c'est le Vecteur Invariant par Permutation (Permutation Invariant Vector ou PIV). L'idée de ce vecteur est de stocker l'information topologique d'une structure donnée, de sorte qu'il suffit de calculer la distance entre deux vecteurs pour avoir la distance entre deux structures [7, 8]. L'avantage de cette variable est qu'elle est très flexible et demande peu de connaissances préalables pour pouvoir étudier des transformations dans un système donné, tout en fournissant des coordonnées de réaction de très bonne qualité, comme cela a été mesuré quantitativement pendant notre étude de la nucléation.

Pour étudier la nucléation, nous avons aussi utilisé l'algorithme Chill+ [9], qui permet de distinguer les différents polymorphes de la glace I en se basant sur les symétries de chaque molécule. Cet algorithme permet de déterminer si une molécule

est dans un état liquide, de glace hexagonale, cubique ou interfaciale. Cette variable collective locale, couplée à des algorithmes de regroupement (*clustering*), va aussi nous permettre de calculer la taille du plus grand noyau pendant le processus de nucléation N_{CHI}. En comptant le nombre de molécules à symétrie hexagonale N_H et celles à symétrie cubique N_C dans le plus grand noyau, on pourra aussi calculer sa cubicité $C = N_C/(N_C + N_H)$, qui mesure simplement le ratio de molécules cubiques et nous informe sur la structure interne du plus grand noyau.

Une dernière variable mérite notre attention. C'est la fonction de réalisation (*committor*). Si nous avons deux états A et B et un système de N particules dans une configuration $x \in \mathbb{R}^{3N}$, la fonction de réalisation $\phi_B(x)$ va mesurer la probabilité du système d'atteindre l'état B avant l'état A. Pour $\phi_B(x) = 0$ nous sommes donc dans l'état A et dans l'état B pour $\phi_B(x) = 1$, et la fonction varie continûment entre 0 et 1 pour des configurations intermédiaires [10, 11]. Concrètement cette fonction nous informe sur la cinétique de la transformation, vu qu'elle nous donne la probabilité de réaliser une transition dans l'état B en partant d'une configuration donnée.

En projetant la fonction de réalisation sur une variable collective quelconque, nous allons pouvoir mesurer sa qualité. Une bonne variable collective déforme de manière minimale le profil de ϕ_B. La méthode d'optimisation du maximum de vraisemblance (*maximum likelihood optimization* [12, 13]) utilise ce résultat pour mesurer de manière quantitative la qualité d'une variable collective.

1.4 Méthodes d'échantillonnage amélioré

Comme dit précédemment, quand une large barrière d'énergie libre sépare deux états métastables, la dynamique moléculaire n'est pas suffisante pour pouvoir échantillonner les propriétés thermodynamiques ou cinétiques d'un système, et on doit la coupler à des méthodes d'échantillonnage amélioré (*enhanced sampling*).

Dans cette thèse, pour obtenir nos résultats finaux nous avons principalement utilisé la méthode d'échantillonnage parabolique (*umbrella sampling* [14]), et deux méthodes d'échantillonnage des chemins de transitions (*transition path sampling* [15]).

L'idée de l'échantillonnage parabolique est de séparer le paysage d'énergie libre en fenêtres qui vont être échantillonnées séparément, pour calculer précisément l'énergie libre. Pour chaque fenêtre, on va ajouter à notre système un potentiel artificiel parabolique, pour contraindre le système à rester dans la zone décrite par la fenêtre. La force de cette méthode est qu'elle est intrinsèquement parallèle, chaque fenêtre étant échantillonnée indépendamment, et qu'elle permet de diminuer énormément le temps d'échantillonnage, pour peu que les fenêtres soit petites et le potentiel suffisamment fort. Nous utiliserons cette méthode dans l'étude de la transition liquide-liquide.

Pour l'échantillonnage des chemins de transition on ne va pas ajouter de biais artificiels, mais on va chercher à construire une chaîne de Markov où chaque pas consistera à générer de manière itérative de courtes trajectoires, que l'on gardera si elles correspondent à une transition entre les deux états étudiés. En itérant suffisamment longtemps, on pourra ainsi générer l'ensemble des trajectoires de transitions. Parmi ses nombreuses variantes, une classe de méthodes consiste à se placer en haut de la barrière d'énergie libre, sur un état de transition, et à tirer des trajectoires depuis cette position pour voir comment elles se détendent. Ce sont deux méthodes issues de cette classe que nous allons utiliser pour étudier la nucléation homogène de

la glace.

1.5 Étude de la transition liquide-liquide

La transition liquide-liquide de l'eau a été initialement postulée pour expliquer la présence de plusieurs formes de glace amorphe à basse température, le polyamorphisme, et certaines anomalies de l'eau, comme sa courbe de densité maximale. Il existe trois formes amorphes : une basse-densité (low density amorhous, LDA) [16], une haute-densité (high density amorhous, HDA) [17] et une très haute-densité (very high density amorphous, VHDA) [18]. Il a été établi expérimentalement que la transition LDA-HDA est de premier ordre [19].

Pour expliquer ce polyamorphisme, l'existence d'une transition liquide-liquide de premier ordre, avec un second point critique dans le domaine sur-refroidi de l'eau, entre un liquide basse-densité (low density liquid, LDL) et un liquide haute-densité (high density liquid, HDL) a été formulé à partir d'évidences numériques [20]. La vérification expérimentale de cette hypothèse est très difficile, car le point critique serait situé dans le *no man's land*, une région où la nucléation homogène de la glace se fait spontanément, prévenant toute observation de ses états liquides métastables.

Néanmoins, de nombreuses études expérimentales ont tenté de sonder les propriétés du no man's land. En général les expériences essayent de limiter la nucléation en utilisant des solutions aqueuses ou des système confinés de taille microscopiques. Des expériences récentes menées sur l'eau salée jettent de forts doutes sur l'existence d'une transition liquide-liquide de premier ordre [21], malgré la mise en évidence d'une transition de premier ordre entre les phases amorphes correspondantes [22], ce qui montre qu'il n'y a pas forcément de lien direct entre les deux.

Depuis la formulation de cette hypothèse de très nombreuses études numériques ont cherché à prouver la présence ou l'absence d'un tel point critique. Les résultats dépendent du modèle utilisé pour décrire l'eau. Pour le modèle ST2, avec tous ses défauts présentés précédemment, la présence d'un point critique a été démontrée sans ambiguïté [23]. Pour le plus réaliste modèle TIP4P/2005, seulement des études indirectes ont été menées, en étudiant les fluctuations de la densité. En utilisant deux théories différentes, mais toujours en se basant sur les fluctuations de la densité et en extrapolant depuis des températures plus élevées, un point critique a été proposé à 182 K et 1.7 kbar [24, 25], puis un autre à 172 K et 1.9 kbar [26].

Toutes ces études sont indirectes, dans le sens où elles ne font pas de calcul d'énergie libre, et ne permettent pas de trancher clairement la question de l'existence ou non d'un point critique pour la transition liquide-liquide du modèle TIP4P/2005. Un des objectifs de cette thèse était de justement calculer rigoureusement les profils d'énergie libre pour trancher cette question.

C'est précisément ce qui a été réalisé dans cette thèse, en utilisant une variable collective S basée sur la PIV pour décrire la transition, couplée avec des calculs poussés d'échantillonnage parabolique. Dans ce cas précis, S est quasiment linéaire à la densité. Nous avons calculé plusieurs points dans l'espace des pressions et températures P, T, sans jamais trouver de barrière d'énergie libre, tous nos profils présentent un seul minimum. De manière remarquable, nous avons pu suivre une ligne de conditions de pressions et de températures où les profils d'énergie libre sont pratiquement plats, ce qui indique que le système peut librement fluctuer dans un large domaine

de densité.

Pour conclure sur la convergence de nos échantillonnages paraboliques, nous avons calculé la fonction d'auto-corrélation de notre variable S dans chaque fenêtre d'échantillonnage. A chaque fois ces fonctions d'auto-corrélation s'annulent sur une durée bien plus courte que celle de nos échantillonnages, ce qui montre leur bonne convergence. De plus, en effectuant une analyse par moyennage de blocs, nous avons estimé nos incertitudes statistiques, en trouvant systématiquement des erreurs inférieures à 1 $k_{\mathrm{B}}T$.

Par ailleurs, à la fois pour confirmer nos profils d'énergie libre et pour obtenir des informations sur la cinétique de notre système, nous avons laissé plusieurs trajectoires se détendre librement, sans biais, d'un état basse ou haute-densité, pour différentes conditions de pression et de température. A chaque fois, les configurations se détendent dans la zone d'énergie libre minimale à plus ou moins 2 $k_{\mathrm{B}}T$, comme attendu. De plus, en reconstruisant les distributions de densité à la fois depuis nos données d'échantillonnage parabolique et depuis les trajectoires libres sans biais ajoutés, nous trouvons des résultats cohérents.

Comme dernière analyse, pour montrer l'absence de barrière d'énergie libre entre les deux états à basse et haute-densité, nous avons étudié le coût de formation d'une interface. En analysant comment les groupes de LDL et de HDL se structurent dans notre système, nous avons pu établir que les deux formes se mélangent de manière quasi-aléatoire, sans chercher à minimiser leur interface.

Nos résultats montrent qu'une métrique basée sur la PIV résout correctement toute la gamme de structures d'eau sur-refroidi pour le vaste domaine P, T exploré. Entre 155 et 182 K et entre 1 et 3 kbar, nous n'avons trouvé aucune preuve d'une séparation entre deux phases liquides, avec la barrière d'énergie libre correspondante.

En particulier, pour des conditions P, T proches de la localisation d'un second point critique (182 K, 1.7 kbar et 180 K, 2 kbar pour comparer avec la Réf. [24], et 170 K, 2 kbar pour comparer avec la Réf. [26]), nous avons toujours trouvé un seul large minimum d'énergie libre, sans barrière. Ce résultat diffère fondamentalement de ce que l'on peut trouver avec le modèle ST2, où une barrière de $4k_{\mathrm{B}}T$ a été observée.

Tous nos résultats montrent donc de façon cohérente l'absence d'un second point critique pour la transition liquide-liquide. Sans diminuer l'importance de tels modèles, qui permettent d'approfondir notre compréhension de phénomènes complexes, cette conclusion montre les limites que peuvent avoir des extrapolations faites à partir d'observation dans d'autre régions du diagramme des phases en utilisant les-dits modèles.

Finalement, il reste à élucider comment le profil d'énergie libre évolue en refroidissant sous la barre des 140 K, où la transition LDL/HDL devient la transition LDA/HDA de premier ordre entre deux formes amorphes. Des calculs rigoureux d'énergie libre seraient encore plus coûteux que pour l'eau sur-refroidie, mais de nos jours ils deviennent accessibles. Il semblerait que la grande bizarrerie et richesse phénoménologique de l'eau, rende inévitable l'utilisation de simulations atomiques détaillées et d'expériences réalisées directement aux conditions de pressions et de températures que l'on veut sonder.

1.6 Étude de la nucléation homogène de la glace

Parmi les nombreux états solides dans lesquels l'eau peut être, les formes polymorphes de la glace I (hexagonale, cubique et en empilement désordonné (*stacking disordered*)) sont les plus intéressants pour nous en tant qu'humain, car ce sont celles que l'on trouve dans les conditions atmosphériques de pression et de température terrestres. C'est pourquoi, malgré le fait que l'eau liquide peut nucléer sous cinq formes de glaces différentes, quand on parle de nucléation de la glace, cela fait généralement référence à la nucléation de la glace I. Ici nous allons principalement nous intéresser à la nucléation homogène de la glace, c'est-à-dire avec de l'eau pure sans aucune impureté ou interfaces qui pourraient servir de site de nucléation et accélérer le processus.

Pour comprendre la nucléation, il est important de bien connaître les polymorphes de la glace I. Elle a deux états bien définis, la glace hexagonale (I_h), qui est la phase stable, et la glace cubique (I_c), qui est métastable. Jusqu'à récemment, aucune observation directe d'un échantillon pure de I_c n'avait été réalisé, et ce n'était même pas clair si elle pouvait être observé dans la nature [27, 28]. Avant ces deux études, la phase observée naturellement était un empilement désordonné métastable, où des couches de glace hexagonale et cubique s'empilent les unes sur les autres, d'où son nom [29, 30, 31]. Comme une caractérisation précise de ce désordre et de son mécanisme de formation est difficile à mener expérimentalement, les études numériques sont de première importance pour comprendre précisément ce phénomène.

Du point de vue de leur symétries cristallines, la glace I_h réside dans le groupe de symétrie hexagonale, la glace I_c dans le groupe de symétrie cubique (d'où leur nom) et la glace I_{sd} dans le groupe de symétrie trigonal. La différence de symétrie entre les glaces cubiques et hexagonales a d'importantes conséquences sur leur empilement. Les quatre faces (111) de la glace I_c peuvent s'empiler sans apparition de défauts cristallin avec la glace I_h , alors que pour cette dernière seulement les deux faces basales (001) peuvent s'empiler avec la glace I_c sans apparition de défauts [31].

Une des théories les plus utilisées pour décrire la nucléation est la théorie de la nucléation classique (*classical nucleation theory*, CNT). Dans cette théorie, les regroupements cristallins de particules (des molécules H_2O dans le cas de l'eau) de n'importe quelle taille sont considérés comme de larges structures cristallines homogènes, avec une fine interface avec le liquide environnant. Dans ce point de vue, où on néglige complètement la structure interne et la forme de ces regroupements cristallins – que l'on va appeler noyau par la suite –, le processus de nucléation peut entièrement être décrit par la différence entre le gain énergétique à former un noyau et le coût énergétique lié à son interface avec la phase liquide. Une fois atteinte une taille critique où les deux coûts s'équilibrent, l'extension du noyau permet au système de réduire son énergie et donc la phase cristalline va s'étendre à tout le système. Ce qui rend la nucléation très compliquée à étudier numériquement, c'est que la formation initiale d'un noyau de taille critique est un événement rare qui nécessiterait en théorie des simulations de l'ordre de la milliseconde – ce qui est inatteignable de nos jours.

L'enjeu de l'étude de la nucléation homogène de la glace est justement de mesurer précisément la taille et la structure interne optimale des noyaux critiques. Les études numériques qui se sont attaquées à ce problème ont principalement utilisé les modèles mW, TIP4P/2005 et TIP4P/Ice. Pour étudier la taille du noyau critique en fonction de la température, plusieurs études ont utilisé la méthode du germe (*seeding*) [32, 33,

34, 35, 36], où on va artificiellement introduire un germe cristallin dans un système liquide, puis générer un grand nombre de trajectoires en comptant celles où il grandit et celles où il rétrécit. Cela permet d'estimer la taille du noyau critique, où on doit avoir à peu près la moitié des germes qui grandissent et l'autre moitié qui rétrécissent. Même si la méthode du germe ne permet pas de faire des estimations très précises, ces études sont très utiles comme point de départ pour étudier la nucléation avec des méthodes plus coûteuses.

En utilisant l'échantillonnage des flux causaux (*forward flux sampling*, FFS) pour reconstruire l'ensemble des chemins de transitions, une étude a pu estimer avec précision la taille du noyau critique $N_c = 474 \pm 12$ et sa cubicité $C = 0.59 \pm 0.07$ à 230 K pour le modèle TIP4P/Ice [37]. Une seconde étude qui utilise la méthode de tir sans objectifs (*aimless shooting*), a reconstruit l'ensemble des chemins de transition, ce qui lui a permis d'estimer la taille du noyau critique $N_c = 450 \pm 35$ et sa cubicité $C = 0.63 \pm 0.05$, à 230 K avec le modèle mW [38]. Une dernière étude a utilisé la métadynamique (*metadynamics* [39]) pour calculer le profil d'énergie libre de la transition, trouvant une barrière d'énergie libre $\Delta G_c = 52 \pm 6\, k_{\mathrm{B}}T$, et estimer la taille du noyau critique et sa cubicité, à 230 K pour le modèle TIP4P/Ice [40]. Cette dernière trouve une taille critique plus faible que les deux études précédentes $N_c = 314 \pm 20$, mais une cubicité similaire $C = 0.7 \pm 0.1$.

L'objectif de notre travail est de produire un ensemble des chemins de transition fiable et de bonne qualité, avec le modèle réaliste TIP4P/Ice, dans la continuité de l'étude réalisée à partir du modèle mW de la Réf. [38]. Cela va nous permettre d'estimer précisément la taille du noyau critique et sa cubicité, avec une analyse quantitative du mécanisme d'empilement désordonné. Mais aussi de mesurer quantitativement la qualité d'une variable collective basé sur la PIV.

C'est précisément ce que nous avons fait à 237 K pour le modèle TIP4P/Ice, en utilisant une méthode rigoureuse d'échantillonnage des chemins de transition. Cette méthode très coûteuse nécessite de générer des milliers de trajectoires de transitions. Deux facteurs rendent ces simulations onéreuses : d'un côté le besoin d'avoir une boîte suffisamment grande pour accueillir un noyau de taille critique (jusqu'à ~ 700 molécules à 237 K, notre boîte étant de 4096 molécules), et de l'autre la longueur minimale des chemins de transitions, $\sim 100ns$ à 237 K.

Pour remplir cet objectif, nous nous sommes appuyés sur deux outils : le premier est une variante de la méthode utilisée dans la Réf [38] : le tir sans objectifs depuis une région donnée ("*aimless shooting within a range*" [41]). Le second est bien évidemment l'utilisation de la PIV pour définir nos variables collectives. Ici nous avons utilisé la distance avec une structure cristalline hexagonale $D_{\mathrm{I_h}}$, qui nous a permis de suivre précisément l'évolution de la structure des noyaux critiques, grâce à une excellente corrélation avec la fonction de réalisation, c'est-à-dire la meilleure coordonnée de réaction.

Ce résultat nous a permis d'offrir une vue détaillée du mécanisme de nucléation, en quantifiant les sources d'apparition d'empilement désordonné. Nos résultats sont cohérents avec les résultats précédemment obtenus sur d'autre potentiels [38] ou avec d'autres méthodes d'échantillonnages [34, 42, 40], et contribuent à clarifier notre compréhension de la nucléation homogène de la glace.

En particulier, nous avons directement observé à la fois l'évolution spontanée d'un germe hexagonal et celle d'un germe cubique, en tirant des conclusions fiables

basées sur la répétition de chaque type de simulation dans 15 répétitions indépendantes. Notamment, nous avons pu observer que l'agrégation de nouvelles molécules à symétrie hexagonale sur le noyau critique était un processus en deux étapes, alors qu'il ne nécessitait qu'une étape pour les molécules à symétrie cubiques. Clairement, nos résultats montrent que la théorie classique de la nucléation est trop simpliste pour pouvoir décrire correctement la nucléation de l'eau.

Un autre résultat de notre étude est de fournir une première mesure quantitative de la qualité de la PIV pour définir des variables collectives, grâce à l'utilisation rigoureuse de la technique d'optimisation du maximum de vraisemblance [12]. En utilisant des données de réalisations massives de la Réf. [38], nous avons trouvé que notre variable $D_{\mathrm{I_h}}$ était la meilleure parmi un jeu de 26 variables collectives, qui était un ensemble très varié, contenant la taille du plus grand noyau, sa forme ou encore l'énergie du système. Il est important de noter que la définition de notre variable est très générale et s'appuie sur très peu de connaissances, puisqu'elle ne nécessite que de définir un état final et initial (ici la glace hexagonale et l'eau liquide). Elle pourrait donc très bien s'appliquer sur d'autre matériaux et à d'autres transformations de la matière condensée.

Pour conclure, le large ensemble de chemins de transitions obtenus avec cette étude va servir de base pour un nouveau projet de recherche. Celui-ci visera à calculer précisément le taux de nucléation, grâce à la reconstruction bayésienne d'un modèle markovien diffusif, comme décrit dans la Réf. [43]. Cette approche s'occupera de ce qui est certainement la question la plus importante dans le large champ de la nucléation homogène : comment calculer directement et avec une bonne fiabilité le taux de nucléation, sans s'appuyer sur les nombreuses approximations de la théorie classique de la nucléation.

1.7 Conclusion

En conclusion, au cours de cette thèse deux problèmes majeurs pour l'eau ont été étudiés : la transition liquide-liquide et la nucléation homogène de la glace.

En utilisant des nouvelles méthodes d'échantillonnage amélioré, couplé avec la nouvelle variable PIV, nous avons pu montrer rigoureusement l'absence de barrière d'énergie libre et de second point critique pour la transition liquide-liquide avec le modèle TIP4P/2005.

Nous avons aussi pu étudier précisément le mécanisme de nucléation homogène de l'eau avec le modèle TIP4P/Ice, montrant que la structure optimale des noyaux critiques est un empilement désordonné, les noyaux purement cubique ou hexagonaux évoluant spontanément vers cette structure. De plus, nous avons pu montrer que la glace hexagonale s'agrégeait majoritairement en deux étapes, là où la glace cubique s'agrégeait en une étape.

Finalement nous avons démontré rigoureusement qu'une variable collective basée sur la PIV permettait de décrire de manière optimale la nucléation.

2 Introduction

2.1 Context

If we look at transformations of condensed matter on the atomic scale, a large number of mechanisms unfold before our eyes. The study of the underlying processes is a long standing problem in the condensed matter community. Due to the evanescent nature of transition states, experimental studies are difficult to carry out and limited in their results. Hence, numerical methods play a major role in our understanding of transformation processes in matter.

Among the many materials that exist on earth, curiously water is at the same time the most ubiquitous and one of the most unusual and challenging to study. Its numerous anomalies at low temperature have been some of the most puzzling to understand in the last decades [1]. Even the fundamental problem of ice nucleation, that is how liquid turn into ice, is not well understood and it still sees a large amount of work being carried out [2].

And yet it is of prime importance to understand transformations and properties of water, as it plays a major role in countless essential domains, ranging from the study of life to climate change (and even in a not so essential domain like aviation). Understanding ice nucleation of water and how it is hampered or favoured by the various substances present in the atmosphere is crucial to understand how clouds form, as they are mainly composed of ice. Hence, it is also a key component to make reliable models about climate change or for meteorology. For biology, many works rely on proper study of proteins folding and unfolding when solvated in water. It is thus of prime importance to have cheap and reliable models to describe water numerically, which in turn imply fine understanding of water and subsequent validation of the models by comparing it with experimental results.

From a more fundamental point of view, the study of water properties has revealed to be an extremely fertile soil for the development of new theories and techniques, whether experimental, analytical or numerical. For instance, the will to understand the supposed liquid-liquid transition of water has led to developing many new experimental techniques for high pressure, ultra-fast x-ray diffraction, confinement, and many ones to study negative pressures [44]. It has also led to development of the theory of thermodynamics of fluid polyamorphism [45].

To put it simply, understanding finely the properties of water, and how to model it, is necessary in basically all of the disciplines of natural science and engineering. Systematic study of the supercooled water properties with general methods is the object of this thesis.

2.2 Thermodynamics and phase transitions

Generally speaking, physical systems have several possible states that they can reach for a given pressure and temperature. For instance, at ambient pressure and summer temperature, water can be liquid, solid or gaseous, but if we take a piece of ice, it will melt into liquid. We say that the liquid state is stable, while the ice state is metastable, and that our piece of ice has performed a phase transition.

To describes this, historically physicists introduced the concept of free energy.

It is a function of thermodynamic quantities, like pressure P and temperature T, and of some order parameter. Without entering too much into the details, an order parameter is a function that describe in which state our system is, telling us for instance if it is liquid or solid. At fixed temperature and pressure, we speak about the Gibbs free energy and will note it G. Metastable states correspond to local minima of G, while the stable state is the absolute minimum of G; different locally stable states are separated by free energy barriers.

In general phase transitions can be put into two group, depending on the continuous properties of G. We speak about first order phase transition when first derivatives of G are discontinuous, like latent heat or density. We speak about second order transition when second derivatives of G are discontinuous, like heat capacity. Despite this relatively abstract definition, one important consequence of the discontinuous nature of first order transitions is that during transition, the system will have some of its parts that have completed the transition and others that haven't, forming an interface between the two that the system will tend to minimize due to its energetic cost. Like for instance when water boils and bubbles of vapor form in it.

The ensemble of equilibrium stability information is generally grouped in a phase diagram, where we show the preferred physical states of matter at different thermodynamic variables. In this diagram we have open spaces where a single state is stable, separated by lines where phase transition will occur, that we call phase boundaries. In other terms, the open space corresponds to free energy with one single global minimum,whereas along the lines we have two or three global minima with the same stability. The most common phase diagram uses pressure and temperature as thermodynamic variables.

If the free energy information contained in a phase diagram is sufficient to describe stability of the states available by a system and the nature of the transition between two states, it has some limitations. As it is an equilibrium quantity, it does not inform us about the out of equilibrium dynamics, like transition pathways and their kinetics. When studying kinetics, one generally wants to reduce the complex high-dimensional phase space dynamics of the system described by the Liouville equation to a more handy and human readable description. This can be done by coarse graining into a few relevant order parameters (or into a discrete set of microstates) and using stochastic equations that represent the average behavior of exact trajectories, to represent how the system passes from one state to another. In this scheme, each minimum of G will effectively trap the system for a time long (from microseconds to millions years) in comparison to typical bond vibrations (few femtoseconds). Rarely, a state will rapidly jump over the barrier into another metastable state.

Studying the kinetics is a more subtle and complex problem than the study of free energy, but a complete reconstruction of kinetics contains more information and is essential to understand the experimental world, where very often the physical systems do not have the time to reach the equilibrium state. A feature that is exploited by organisms to maintain their living state, and also by humans for the synthesis of complex materials for technological applications. In this thesis, we will address both the thermodynamics and kinetics of transformation processes in water.

2.3 Numerical study of transformations of matter

As briefly stated previously, transformation processes are hard to grasp experimentally, due to the small time and length scale involved. In theory, molecular dynamics simulations would be ideal to study such phenomena at microscopic scale, since long trajectories of the system would sample fluctuations within metastable states and all their transitions among them. In practice however, there are two main limitations.

The first one is linked to the way we describe our material numerically. A quantum approach is a priori more precise as it relies on first principles quantum mechanics, but it drastically limits the length scale accessible during a simulation due to computational cost. Thus it is in general of limited interest when we study phase transitions, as if size effect are important for the transition (and they generally are), the limited size of the system under study would spoil the meaning of the extracted data. Thus classical approaches are generally used when studying transformations of matter, as they give access to length scales thousands of times larger than in quantum simulation. The problem is that now we need to represent our material with some predefined force fields, generally tuned to reproduce at best properties found with either quantum simulation or experimental data. As you may guess, this approach can only give approximate results, even if some models reproduce quite well selected properties of the material under study.

The second limitation come from the current time scale available by computation, as it is orders of magnitude smaller than what we would need for proper sampling of transitions in condensed matter. Quantum simulations are limited to the sub-nanosecond time scale, and classical atomistic simulations are limited to the sub-milliseconds time scale. (For instance ice nucleation might require from millisecond to millions years depending on conditions). This limits drastically the range of transformations that one can study. To solve this problem and also to get insight not automatically provided by a simple long trajectory, several methods were developed in the last decades, called enhanced sampling methods [46].

2.3.1 Enhanced sampling methods

Despite being grouped under one terminology, enhanced sampling methods tend to solve two different problems : accelerated exploration of the available metastable states and/or precise sampling of the thermodynamic and kinetic properties. The latter consists in large accumulation of samples in relevant regions of the configuration space, to construct an estimate of equilibrium probability distributions (and possibly kinetic properties). In the present thesis, we used enhanced sampling methods that exploited the two following ideas :

1. add artificial, external biasing force to the natural forces of the system, in a way that enhances population of barrier regions, compared to its negligible equilibrium value.
2. generate several trajectories starting from specific configurations, only keeping those that fulfill some clever requirements.

Even if the methods are general and do not restrain themselves to the study of a specific material, what limits their use is the need to define some variables that

will allow us to track the transformations studied. Such variables are generally called structural descriptors, collective variables, order parameters, or reaction coordinates by different scientific communities. We will talk about reaction coordinates or order parameters when their aim is to describe a specific reaction mechanism.

2.3.2 Reaction coordinates

Defining a reaction coordinate is not a difficult task, but it can be extremely challenging to define a good reaction coordinate. Often this relies on specific knowledge about the system under study and tedious trial-and-error iterative process. Postponing the discussion about methods to quantitatively assess the quality of a variable, it is important to note that efficiency of enhanced sampling methods is directly linked to the quality of the reaction coordinates. Moreover, finding optimal reaction coordinates can lead to deeper understanding of the transformation processes.

A recently developed approach to define good reaction coordinates, regardless of the system under study, starts by describing transformations of matter as changes in a matrix that contains the inter-atomic bond network information [47, 7, 8]. One aim of this thesis is to apply, improve and further validate this approach by applying it to the study of water in the low temperature regime.

2.4 Complexity of water

Despite its molecular simplicity, water reveals to be very complex from the viewpoint of physico-chemical properties. These include numerous triple points, at least one critical point, more than 10 stable solid phases, no less than 3 metastable amorphous phases and several theoretical liquid states [48], without speaking of its 74 anomalies compared to other liquids ! Among them, one of the most famous is that water expands upon freezing, increasing its volume by 9% under atmospheric pressure, implying that ice floats on water. Another one is its high density, with a maximum at 4°, meaning that upon cooling or heating liquid water its density will decrease. This is why for instance the bottom of fresh water lakes keeps the same temperature regardless of the external temperature variation, as water at 4° will sink due to its higher density. This temperature-driven shift in density is also the origin of ocean currents such as the gulf's stream, which have huge impact on land climate.

Just to have a glimpse of its complexity, figure 2.4.1 presents the stable phase diagram of water [48]. Every solid line represents a phase boundary. Along them, two phases will stably coexist in any relative proportions. When three such lines join, we have a triple point with three stably coexisting phases. As already mentioned, it is important to note that such diagram indicates the equilibrium properties of water, and does not tell us about how the phases are kinetically connected, or by which microscopic mechanisms water transforms from one phase to another when crossing a line.

The complexity of collective behavior of water is linked to its polarized molecule H_2O. This polarization is due to the difference in electronegativity between oxygen and hydrogen nucleus, which leads to the hydrogen's electron being more attracted by the oxygen atom, resulting in a slightly negatively charged oxygen atom, while hydrogen is slightly positively charged. As a result water molecules will form hydrogen bonds, where one of the hydrogen atom is linked to the oxygen atom of another

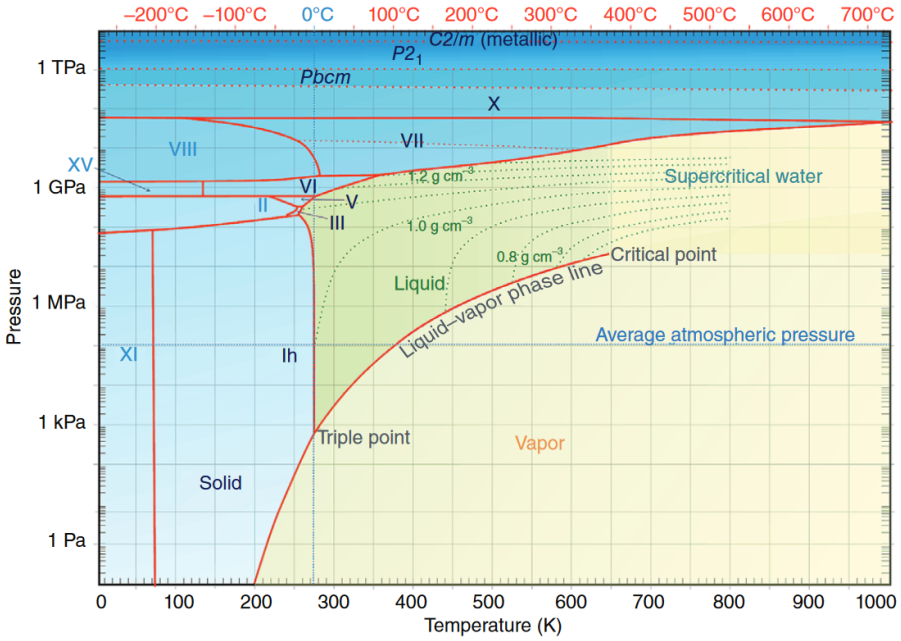


FIGURE 2.4.1 – Stable phase diagram of water

molecule. Compared to other types of chemical bonds, the hydrogen bond is neither strong nor weak, meaning that it can be easily broken but will generally survive to thermal fluctuations [49].

At moderate pressures, each water molecule can form 4 hydrogen bonds, 2 involving its oxygen atom and 2 involving its hydrogen atoms. This 5 water molecules bonded together will optimally arrange themselves in tetrahedral shapes as shown in figure 2.4.2. In solid phases this local tetrahedral arrangement will extend to the whole system and produce crystalline structures. In liquid phases thermal energy will break, stretch or bend the hydrogen bonds, leading to only local clusters of tetrahedral structure, even if large chains of hydrogen-bonded molecules are present. At high pressure this tetrahedral structure will be enriched by supplementary water molecules and the soft hydrogen bond will evolved toward a strong shared proton due to nuclear quantum effect [50].

Properly describing water with classical force fields is not an easy task. As a result, until quite recently they was no model able to reproduce correctly water physical properties, with the correct phase diagram and the correct structural properties of each phase [5].

Here, we will restrict ourselves to the study of water at low temperature, from 140 to 237 K, and moderate pressure, from 1 to 5000 atmospheres (0.1 MPa to 0.5 GPa). In this region, two transformations are of main interest : the supposed liquid-liquid transition of water and the homogeneous ice nucleation of water.

2.4.1 The “liquid-liquid transition”

If we look at water at very low temperature (less than 140 K), we see that water possesses at least three amorphous metastable phases : a low density amorphous (LDA), a high density amorphous (HDA) and a very high density amorphous (VHDA). This property of a solid to exist under different amorphous form is called

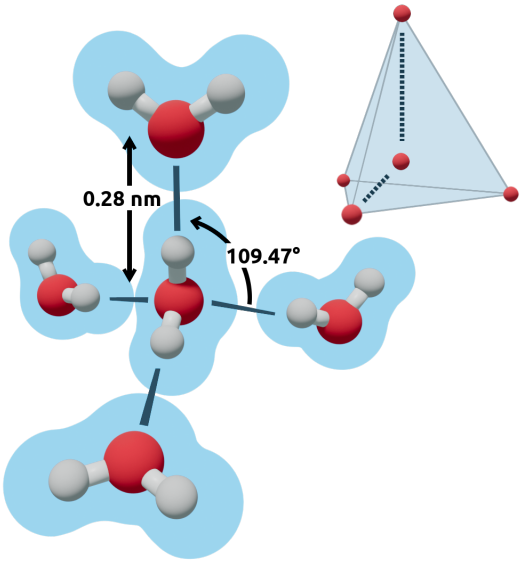


FIGURE 2.4.2 – Schematic representation of a tetrahedral hydrogen-bonded water pentamer

polyamorphism. Evidence of these three forms cames from experiments [17, 16, 18]. The low and high density amorphous phases are connected by a reversible first-order transition [19].

To explain the existence of these states and some of the water anomalies, several theories were formulated in the last three decades. Among them, one postulates the existence of two structurally different liquid, with a first order transition between them [20]. This hypothesis is hard to be verified experimentally in pure water. Indeed, its supposed location is right in an area, the so called *no man's land*, where liquid water spontaneously transforms into ice, preventing any experimental measure on the underlying metastable liquid states. Its numerical study is also not trivial to pursue, as one needs to distinguish two similar liquid structures and to perform simulation at low temperature with slow dynamics, that renders very costly the statistically meaningful sampling of the corresponding states.

In fact, an uninterrupted series of experimental works have tried to probe the existence or absence of this transition, using salty water or confined systems to prevent freezing [51, 52, 53, 54, 55]. Results show that in salty water, despite existence of a LDA-HDA first order transition, a liquid-liquid transition does not exist [22, 21]. But for pure water the nature of the no man's land remains unobserved.

In parallel, the computational scientific community has not been at rest, with an ongoing debate about the very existence of the liquid-liquid transition. Most of the debate crystallized around the sampling methods *and* the model used to describe water during the simulation. After long controversies, a study based on a sub-optimal model of water found clear and indubitable evidence of a first order liquid-liquid transition [23]. Whilst other studies found indirect evidence that this transition might not exist using a more recent and precise model of water [56, 57].

Most likely the debate in the scientific community will continue until experiments bring it to an end. Still, one of the aims of this thesis is to clarify the situation with robust evidence of the existence or absence of the transition with the use of a state-

of-the-art water model and of enhanced sampling methods.

2.4.2 Homogeneous ice nucleation

Everyone knows that water freezes when it comes down to 273,15 K (0° C), but few know that in pure water the kinetic barrier is so large that freezing would require more than the age of the universe [36]. In practical situations, the freezing is accelerated by the presence of impurities or interfaces, that favor the apparition of small nuclei of ice subsequently growing until filling the whole system once they have attained some critical size. In the case of perfectly pure water we speak about homogeneous ice nucleation, otherwise we speak about heterogeneous ice nucleation. To study the mechanism of homogeneous ice nucleation, experiments are limited by the difficulty to prepare pure water samples and by the short time and length scale involved in the microscopic nucleation processes. One may ask what are the structural properties of the initial nucleus of ice, but for instance at 230 K, its size is only of a few nanometers and its growth can take place in few hundreds of nanoseconds. Note that if the growth of the nucleus is fast, its apparition is a rare event, so one may need to wait milliseconds or seconds to see one appear and propagate to the whole system.

We want to stress that despite the experimental difficulties in realizing homogeneous nucleation of ice, the process remains of fundamental interest to understand crystallization and to address the more complex case of heterogeneous ice nucleation. The latter is more readily observed in real life, however it has additional difficulties related to the precise experimental characterization of the nucleation sites (defects, impurities, etc.). Also from a numerical viewpoint, further complications arise from the definition of interactions between different molecular or atomic species. It is a very challenging domain, where theoretical predictions are still scarce and often in strong disagreement with experiments. This is also linked to the limitation of the simplified models used, like classical nucleation theory.

If the time and length scales of detailed nucleation mechanisms are too small for experiments, they are on the reverse too big for simple numerical study. In fact, increasing the number of molecules increases rapidly the cost of the simulation, so either we can use less precise simulations or we can resort to smaller number of molecules. But even by limiting the number of molecules, milliseconds or seconds are just too much for nowadays supercomputer. So we need to use enhanced sampling methods, with all the complications around the definition of a reaction coordinate implied.

One of the aims of this thesis is to study nucleation with state-of-the-art molecular models and enhanced sampling methods, coupled with the recent general approach to compute reaction coordinates based on the adjacency matrix.

2.5 Open challenges

As discussed above, on the one hand we need to develop general tools and methods to study transformations of matter. On the other hand water is a material that despite its molecular simplicity gives rise to highly complex behaviors, and requires special care to be investigated, especially in the supercooled region.

Can we successfully develop general methods to study transformations of matter ? And can we apply them to the difficult study of both the liquid-liquid transition and homogeneous ice nucleation of water ?

For the liquid-liquid transition, we need to study water at several low temperature conditions, to see how the stability of the hypothesized high and low density liquid states evolves and in which way they are mixing or coexisting together. In particular, a recurring question in the community concerns the reconstruction of free energy landscapes using accurate inter-atomic potentials, which is one of the aims of our work.

For homogeneous ice nucleation, we need to study the critical nuclei and whether and how they spontaneously evolve from a "pure" crystalline state with unique symmetry to a disordered stack of hexagonal and cubic ice layers. We remark that the mechanism is believed to be far from trivial, with a complex interplay of different crystalline phases, and to defy classical nucleation theory.

To achieve these targets, we realised massive classical molecular dynamics simulations (more than 10 millions of cpu hours) coupled with a range of different enhanced sampling methods, to systematically reconstruct the mechanisms, the free energy landscapes and the kinetics of several transformation processes in water.

We anticipate that for the first time we reconstructed accurate free energy landscapes in no man's land to study the liquid-liquid transition with the TIP4P/2005 potential, observing the lack of free energy barriers and of a discontinuous transition, similarly to mW and contrary to the ST2 potential. We also provide robust evidence of the nucleation mechanism, starting from hexagonal or cubic nuclei that spontaneously evolve toward stacking disordered nuclei, and eventually ice, applying for the first time rigorous aimless shooting techniques to the accurate TIP4P/ice potential. Our results provide also a solid background and database for further investigations of nucleation free energy landscapes and kinetic rates.

This thesis is structured in two parts. The first one presents the various methods and numerical tools we used, with their underlying physical and mathematical principles. The second one presents results obtained during the study of the liquid-liquid transition, or of the homogeneous nucleation of ice.

Première partie

Methods

3 Simulations of Condensed Matter

The study of the transformation processes at an atomic scales is a long standing problem in material science. The short time and length scale of these transformations makes them difficult to study experimentally. This is why numerical simulation if of prime importance to get knowledge over these phenomena, as it gives us access to a whole new set of time and length scale, that are often better fitted to the problem we want to solve. Of the various general methods available, molecular dynamics simulations are well suited to study thermodynamic and kinetic properties for a given set of temperature and pressure. One can play with either classical, based on force fields, or *ab inito*, based on density functional and quantum mechanics, molecular dynamics. In this thesis we will only use classical molecular dynamics and refer to it simply as molecular dynamics or MD when lazy.

It is interesting to think about molecular dynamics simulations *as* experiments where we have full access to microscopic data of the system. However, we should never though that they *are* experiments, as we are always limited by the precision of the force fields used to describe our material and to the available time scale for nowadays computers.

3.1 Classical molecular dynamics

3.1.1 Short introduction to statistical principle

To better understand the idea behind molecular dynamics, its necessary to do a brief recap of the underlying statistical principles.

Phase space

When we study the dynamic of a system compound of particles using classical mechanics, to reconstruct a particle's trajectory we need its position $\mathbf{r}(t)$ and momentum $\mathbf{p}(t)$ at each time. For a time t and one particle, the state of a 3 dimensional system are thus defined by the set of all coordinates $x(t)$, $y(t)$, $z(t)$, $p_x(t)$, $p_y(t)$ and $p_z(t)$. This form an abstract 6 dimensional space that we call the phase space of the system.

If we take a larger number of particles, say N particles, we get a $6N$ dimensional space. The dynamic of the system is then described by a trajectory in this abstract space. Each point of the phase space describe a microscopical state of the system. The macroscopic equilibrium state result of an average over the microscopical states accessible by the system, governed by a statistical law.

Ergodicity

To describe a macroscopic state, we can either take a large number of system with the same initial conditions and measure their states, or take one system and measure its states over a long period of time.

Let's say that we want to measure a physical quantity A. If we are doing a large amount of measure over time, we could compute its temporal average

$$\langle A \rangle = \sum A_k p_k \tag{3.1.1}$$

where p_k is the probability to be in a microscopic state A_k over time.

But if we are doing a large amount of measure over equivalent system, we could compute the ensemble average

$$\overline{A} = \frac{1}{N} \sum_k A_k N_k \tag{3.1.2}$$

where N_k is the number of time we have measured the microscopical state A_k. When N become very large we have

$$\lim_{N \to +\infty} \frac{N_k}{N} = p_k \tag{3.1.3}$$

Which in return implies that the two average become equals

$$\langle A \rangle = \overline{A} \tag{3.1.4}$$

And reciprocally, this means that average over a long period of time is equal to ensemble average. This is the ergodicity principle.

It is of first importance, as it means that by simulating long enough trajectory, we will be able to reconstruct macroscopic properties of any system. And this is precisely the idea of molecular dynamics : to compute time evolution of a set of particles to sample their phase space over a long period of time and thus to reconstruct the particles macroscopic properties.

3.1.2 Time evolution

Time evolution of a system is governed by the simple principle of action minimization, that allow us to compute equation of motion for any system. Even though we are generally not able to solve them analytically and need to use approximate numerical methods.

The NVE ensemble

If we consider an isolated system of N particles of mass m in a box of volume V, we have a Lagrangian composed of two terms

$$\mathcal{L} = \sum_{i=1}^{N} \frac{1}{2} m_i \mathbf{v}_i^2 - U(\mathbf{r}_i) \tag{3.1.5}$$

the first one represent the kinetic energy of our system and the second term represent the potential energy due to interaction between the particles. In this set-up, N, V and the total energy E of the system are constants (as the system is isolated). This is why we call this the NVE ensemble.

We can go from Lagrangian to Hamiltonian formulation using a Legendre transformation $\mathcal{L}(\mathbf{r}, \mathbf{v}, t) \longrightarrow H(\mathbf{r}, \mathbf{p}, t)$, defining the momentum as $\mathbf{p}_k = \partial_{\mathbf{v}_k} \mathcal{L}$ and the Hamiltonian H as

$$H = \sum_{i=1}^{N} \mathbf{p}_i \left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_i} \right) - \mathcal{L} \tag{3.1.6}$$

From the Lagrangian, using the least action principle, we can deduce the Euler-Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{v}_i}\right) = -\frac{\partial \mathcal{L}}{\partial \mathbf{r}_i} \tag{3.1.7}$$

from which we can compute the equation of motion

$$m_i \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = -\frac{\partial U}{\partial \mathbf{r}_i} \qquad \text{and} \qquad -\frac{\partial U}{\partial \mathbf{r}_i} = \mathbf{F}_i \tag{3.1.8}$$

$\mathbf{F}_i$ is the force that act on the particle i. Except for very simple form of potential, it is not possible to solve analytically this set of differential equations. Thus we should use numerical methods to resolve these equation of motion.

To do that, we will discretize time using a time step Δt. It should be small enough so that the force could be considered as constant between two time steps. At each iteration we compute the force, then using equation of motion we update velocities and positions.

Verlet and leap-frop algorithm

The two most common algorithm used to integrate equation of motion in molecular dynamics are the verlet and leap-frog one. They are both equivalent, and have the right properties for integrator : they conserve the phase space volume, the energy and total momentum, and they are time reversible. The first properties is necessary to achieve ergodicity [58].

The two algorithm rely on a second order Taylor expansion of the position (in the following we will omit to specify that they act on the i-th particles for simplicity)

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \frac{\mathrm{d}\mathbf{r}(t)}{\mathrm{d}t}\Delta t + \frac{1}{2}\frac{\mathrm{d}^2\mathbf{r}(t)}{\mathrm{d}t^2}\Delta t^2 + O(\Delta t^3) \tag{3.1.9}$$

If the take the same expansion for $r(t-\Delta t)$ and sum the two, we obtain

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \frac{\mathrm{d}^2\mathbf{r}(t)}{\mathrm{d}t^2}\Delta t^2 + O(\Delta t^3) \tag{3.1.10}$$

Now we have two ways to define the velocity, that will either lend us the verlet or leap-frog scheme

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} + O(\Delta t^2) \qquad \text{(verlet)} \tag{3.1.11}$$

$$\mathbf{v}\left(t - \frac{1}{2}\Delta t\right) = \frac{\mathbf{r}(t) - \mathbf{r}(t-\Delta t)}{\Delta t} + O(\Delta t^2) \qquad (\text{leap} - \text{frog}) \tag{3.1.12}$$

Using the leap-frog scheme and the equation of motion (3.1.8), we obtain the following relations

$$\mathbf{v}\left(t + \frac{1}{2}\Delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\Delta t\right) + \frac{1}{m}\mathbf{F}(t)\Delta t + O(\Delta t^2) \tag{3.1.13}$$

$$\mathbf{r}\left(t + \Delta t\right) = \mathbf{r}(t) + \mathbf{v}\left(t + \frac{1}{2}\Delta t\right)\Delta t + O(\Delta t^2) \tag{3.1.14}$$

The position of the particle i is computed at time $t + \Delta t$ from its velocity at time $t + 1/2\Delta t$, just like two frog leaping over each other back (with a bit of imagination..).

3.1.3 Temperature coupling

As its fine to compute things in the NVE ensemble, it doesn't allow us to control precisely the temperature and pressure. To compare the simulation result with experiments hence become a bit tricky, as they are generally done at fixed pressure and temperature condition.

To fix temperature, we couple our system with a thermostat of constant temperature T. In this configuration, the number of particles N and volume V are still fixed, but now the energy can freely evolve and its the temperature that is constant. So we call this one the NVT ensemble.

Generally speaking, the temperature of a system is fixed by its kinetic energy, such that, for a system of N particles, with masses $\mathbf{m}_i$ and velocities $\mathbf{v}_i$

$$E_k = \frac{1}{2}\sum_{i=1}^{N} m_i \mathbf{v}_i^2 \tag{3.1.15}$$

$$\frac{1}{2} N_{df} k_{\mathrm{B}} T = E_k \tag{3.1.16}$$

where k_{B} is the Boltzmann constant and N_{df} is the number of degrees of freedom. It can be computed from the number of constraints N_c imposed on the system and the number of translational and rotational degree of freedom accessible by the center-of-mass N_{com} :

$$N_{df} = 3N - N_c - N_{com} \tag{3.1.17}$$

To fix the temperature they are now three main methods that can be used.

Nosé-Hoover thermostat

The first one was introduced by Nosé [59] and then enhanced by Hoover [60]. In this scheme, we add new terms to our Lagrangian to represent a thermal reservoir and a friction force. This force is proportional to the product of each particle's velocity and a heat bath parameter η which posses its own velocity v_η and "mass" Q.

The equation of motion of the particles gain a new "thermic" term :

$$\frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = \frac{\mathbf{F}_i}{m_i} - Q v_\eta \mathbf{v}_i \tag{3.1.18}$$

where the equation of motion of the heat bath parameter is dependent of the fixed temperature T_0 and the current temperature T :

$$\frac{\mathrm{d}v_\eta}{\mathrm{d}t} = (T - T_0) \tag{3.1.19}$$

However this simple thermostat can exhibit non-ergodic behavior for low dimensionality system. This can be corrected by introducing chain of thermostats that improve its ergodicity. Even if it is still not perfect [61], it yields a correct NVT ensemble [62].

Berendsen thermostat

The principle of the Berendsen algorithm is to suppress the fluctuation of the kinetic energy. This yield an improper NVT ensemble and so technically the sampling is not correct. But the error scales like $1/N$, so its not necessarily a big deal for very large systems. Except for the distribution of kinetic energy and fluctuation properties that are obviously not correct [63].

The velocities of each particle are scaled every step with a time dependent factor λ such that the kinetic energy is scaled at each step by

$$\Delta E_k = (\lambda - 1)E_k \tag{3.1.20}$$

And the λ factor itself is given by

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T(t - 1/2\Delta t)} - 1\right)\right]^{1/2} \tag{3.1.21}$$

where the parameter τ_T is not exactly equal to the time constant τ of the temperature coupling

$$\tau = 2C_v \frac{\tau_T}{N_{df} k_{\mathrm{B}}} \tag{3.1.22}$$

with C_v the total heat capacity of the system, k_{B} the Boltzmann's constant and N_{df} the total number of degree of freedom. The reason of this inequality between τ_T and τ is that when we rescale the velocity, the energy difference is distributed between kinetic and potential energy. Hence a smaller scaling in temperature than in energy [64].

With this scheme, the deviation of the system temperature T from the thermostat temperature T_0 is slowly corrected according to

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau}. \tag{3.1.23}$$

This means that the temperature deviation decay exponentially with a time constant τ. So we can fix freely the strength of the coupling and its influence on the conservative dynamics, by taking short coupling time (like 0.05 ps) or long coupling time (like 2 ps).

Velocity-rescale thermostat

The velocity-rescale thermostat is basically the same as the Berendsen thermostat, but with an additional random term [65]. This term ensures a correct energy distribution by modifying it according to

$$\mathrm{d}E_k = (E_k^0 - E_k)\frac{\mathrm{d}t}{\tau_T} + 2\left(\frac{E_k E_k^0}{N_{df}}\right)^{1/2} \frac{\mathrm{d}W}{\sqrt{\tau_T}} \tag{3.1.24}$$

with $\mathrm{d}W$ a Wiener process, N_{df} the number of degree of freedom, E_k the kinetic energy and E_k^0 the thermostat kinetic energy. A Wiener process is a random process which is continuous, with gaussian distribution and no memory. This scheme produces a correct NVT ensemble [62] and has all the advantages of the Berendsen thermostat : first order decay of temperature deviations and no oscillations in the decay. Plus the finely tune-able coupling time.

3.1.4 Pressure coupling

In the same spirit as for the temperature, to fix the pressure we will couple our system with a barostat of constant pressure P in addition to our thermostat. In this configuration, the number of particles N is still fixed, but now the volume and energy can freely evolve and its the temperature and pressure that are constant. We call this one the NPT ensemble.

The pressure of a system is fixed by the difference in energy between the kinetic energy and the internal pair potential (the virial)

$$P = \frac{2}{3V}\left(E_k - \Xi\right) \tag{3.1.25}$$

with V the volume and Ξ the virial of the system, define as such

$$\Xi = -\frac{1}{2}\sum_{i<j} \mathbf{F}_{ij}(\mathbf{r}_i - \mathbf{r}_j) \tag{3.1.26}$$

where $\mathbf{F}_{ij}$ is the force on particles i due to the particle j.

Berendsen barostat

Following the definition of the pressure (3.1.25), one way to change the pressure is to modify the virial. The Berendsen algorithm does just that, by scaling interparticles distances every steps [66].

In general the scaling will be anisotropic and give rise to a scaling matrix μ such that

$$\mu_{ij} = \left[\delta_{ij} - \frac{\Delta t}{3\tau_p}\beta_{ij}\left(P^0_{ij} - P_{ij}(t)\right)\right]^{1/3} \tag{3.1.27}$$

where β is the isothermal compressibility of the system and $\mathbf{P}^0$ the fixed pressure of the barostat. As the equation of motion are modified by pressure coupling, the conserved energy also needs to be corrected by the work the barostats applies to the system. This way of fixing the pressure can lead to large oscillations of pressure and volume.

This scheme has the same effect of a first order relaxation of the pressure towards the given reference pressure $\mathbf{P}^0$, according to

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = \frac{\mathbf{P}^0 - \mathbf{P}}{\tau_p} \tag{3.1.28}$$

and so it has the same flexibility as the Berendsen algorithm for thermal coupling, as one can choose short or long time of relaxation.

Its important to note that by construction, as we are rescaling positions and not adding term to the hamiltonian, this barostat does not yield a true NPT ensemble. Even if the average pressure will be rightly fixed, other physical quantities like volume or enthalpy may be totally off in comparison to a true NPT ensemble. So it is required to use more accurate barostat when this quantities need to be precisely evaluated, like the Parinello-Rahman one presented just below [62].

Parinello-Rahman barostat

A more precise and correct scheme to fix the temperature is the Parinello-Rahman barostat, which is similar to the Nosé-Hoover temperature coupling presented above. This give rise to a new hamiltonian and a correct NPT ensemble [67] [68]

$$H = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}_i) + \sum_i P_{ii} V + \frac{1}{2} \sum_{ij} W_{ij} \left(\frac{\mathrm{d}b_{ij}}{\mathrm{d}t} \right)^2 \tag{3.1.29}$$

with the following equation of motion

$$\frac{\mathrm{d}^2 \mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathbf{F}_i}{m_i} - \mathbf{M} \frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} \tag{3.1.30}$$

$$\mathbf{M} = \mathbf{b}^{-1} \left(\mathbf{b} \frac{\mathrm{d}\mathbf{b}'}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{b}}{\mathrm{d}t} \mathbf{b}' \right) \mathbf{b}'^{-1} \tag{3.1.31}$$

where $\mathbf{W}$ is a matrix parameter that determines the strength of the coupling, $\mathbf{b}$ is the box vector represented as a matrix and V the volume of the box.

The box vector $\mathbf{b}$ obey the following equation of motion

$$\frac{\mathrm{d}^2 \mathbf{b}}{\mathrm{d}t^2} = \mathbf{V} \mathbf{W}^{-1} \mathbf{b}'^{-1} \left(\mathbf{P} - \mathbf{P}^0 \right) \tag{3.1.32}$$

So contrary to the Berendsen barostat, this one will exhibit oscillation during relaxation toward the reference pressure. If the system current pressure is far from the equilibrium pressure, this will result in large box oscillation that may crash the simulation. This is why, despite its better precision and correctness, it is often necessary to use Berendsen pressure coupling for a first equilibration before resorting to this one.

3.1.5 Periodic boundary conditions

Simulations are performed on system with finite size, which we call simulations box. In classical simulations, their sizes are generally on the scale of nanometer, containing hundreds or thousands of particles. To have an order of magnitude in head, a 1 cm^3 piece of matter contains roughly $\sim 10^{23}$ particles. And if we look at the ratio of surface over volume, it would be 1 millions time higher for a simulated cubic box of 10 nm than for a real sized system of 1 cm. This means that all computed properties would be spoiled by appearance of unwanted surface interactions.

To circumvents these limitations, it is necessary to introduce periodic boundary conditions. In this scheme, the simulation box is replicated in all directions, in the same way one would build pavement on the ground, but in three dimensions. If we have a cubic box of size $\mathbf{L} = (L_x, L_y, L_z)$, an atom located on $\mathbf{r} = (x, y, z)$ would have a fictitious "image" atom in $\mathbf{r} = (x + n_x L_x, y + n_y L_y, z + n_z L_z)$, where $\mathbf{n} = (n_x, n_y, n_z) \in \mathbb{N}^3$.

In fact, this scheme take care of the unwanted artifact due to edge effect, but as you may guess it introduces its on set of artifacts. In crystalline system, periodic boundary conditions are desired and doesn't cause much harms. But in non-periodic system like liquids, the periodicity will causes errors due to the un-physical nature of the replication. Those are less severe that edge effect, and will be reduced as the box size is increased.

Long and short-range summation

As already mentioned, particles interaction are encoded in a potential energy terms, which is a function of all the atoms positions $U(\mathbf{r})$. Theoretically, this potential can be decomposed into terms that implies interaction of pairs, triplet, up to n-uplets of particles. In practice however, potential is often modeled with terms that implies only pair and triplet terms. These pair and triplet terms can be further divided in two categories : the short and long-range interactions. We will classify those that decrease faster than $1/r^3$ as short-range, and all others as long-range.

For short-range terms, one can define a cutoff radius over which the interaction is equal to zero and so no computations are requested when two particles are separated by more than this radius, allowing huge performance gain. Also the minimum image convention is used : only nearest particles will be considered, either it being image or real. This implies that the cut-off radius used to truncate short-range interactions may not exceed half of the shortest box vector, or otherwise there would be more than one image within the cut-off distance of the potential.

Long-range terms cannot be cut in this way and should be taken in their entirety. The most used technique to achieve this is the Ewald summation scheme [69]. For long-range interaction we have an infinite sum composed of one term with slow convergence. The idea is to decomposed this sum into two terms with quick convergences and a constant term

$$E_{long-range} = \sum_{\mathbf{n}} \sum_{i,j}^{N} \phi(\mathbf{r}_{ij,(n)}) = E_{dir} + E_{rec} + E_0 \tag{3.1.33}$$

where ϕ is some long-range term, E_{dir} is a sum in real space which contains screened short range interactions, and E_{rec} is a sum in reciprocal space which contains the long range interactions. Real and reciprocal space are connected by Fourier transform. This decomposition allow one to use small cut-off in direct space, of the order of less than 1 nm for the direct part.

The reciprocal sum is still a problem in term of performance, as it scale as the square of the number of particles N^2, making it not fit for large system. Fortunately, the particle-mesh Ewald method was invented to improve performance of the reciprocal term [70]. In this scheme, the charge are assigned to a grid. The grid is then Fourier transformed and the reciprocal energy term is obtained by a single sum over the grid. The potential at the grid points is then calculated by inverse transformation to retrieve the forces on each atom. This algorithm scales as $N \log(N)$, making it more fit to large system than the simple Ewald summation technique.

3.2 Force fields

We spoke about all the complex ways to deals with short and long-range interactions in molecular dynamics, but we have not yet spoken about how we define the inter-atomic interactions. To do that one will use models, or force fields, specifically defined to describe the molecules under study.

Force fields are generally separated into two descriptive part. One that describes the molecule's geometry with distances and angles between particles. And one that

describes how the molecules will interact on short and long-range with others molecules, using an interaction potential. In general this potential is a combination of electrostatic interaction with a Lennard-Jones potential, as for ST2 and TIP4P-like model. The Lennard-Jones potential has a strong repulsive behavior on short-range and a weak attractive behavior on long-range. Thus it effectively encode in a simplified way the repulsion due to overlapping of electron orbitals and small forces linked to small polarization of molecules. But some model can also be composed of purely short-range terms, as for the mW model for instance.

Models aim to achieve the best fit between right estimations of the molecules properties and computational efficiency of the model. Depending on the context, some prefer to sacrifice a bit of computational speed, while other prefer to sacrifice physical precision. They are often constructed in two step :

— first a specific potential energy choice is made to describes the molecule, with a set of parameters $\{\lambda_1, \ldots, \lambda_n\}$ that can be adjusted to change shape and behaviors of the potential

— second the set $\{\lambda_1, \ldots, \lambda_n\}$ is fitted to reproduce a specific physical properties or a set of physical properties. It can also be adjusted to fit energies estimated with simulation based on quantum principle.

After that the model quality is evaluated by how it reproduces non-fitted physical properties.

Here we will present some of the dozens of models that exist to describe water, with their physical and computational limitations. Among them, we only used TIP4P/Ice and TIP4P/2005 during our simulations, but we often compared results obtained with ST2 and mW, so it's good to have a small overview of them.

3.2.1 The mW model

Contrary to most models of water, the monoatomic water model mW is a coarse grained one, meaning that water will be represented by a sole particle. Despite this simplified pictures, which grants the models great computational performances, mW reproduce quite well various properties of liquid water, like density, and its ice I structures [3].

Its success is based on a clever observation. Silicon and water, despite their lack of chemical similarities, behave in the same way when we look at their physical properties. The only feats they have in common is that both form tetrahedrally coordinated structures. Hence mW was defined upon a silicon's model, by tuning some coefficients to further favors tetrahedral structures.

Precisely, the potential energy of the model is a function of the inter-atomic pair distances and the angles formed by triplet of atoms, defined as

$$E_{mW} = \sum_{i,j>i} \phi_2(r_{ij}) + \sum_{i,j\neq i,k>j} \phi_3(r_{ij}, r_{ik}, \theta_{ijk}) \tag{3.2.1}$$

$$\phi_2(r) = AB\epsilon \left(\frac{\sigma}{r}\right)^4 \mathrm{e}^{\gamma\sigma/(r-a\sigma)} \tag{3.2.2}$$

$$\phi_3(r_1, r_2, \theta) = \lambda\epsilon \left(\cos\theta - \cos\theta_0\right)^2 \mathrm{e}^{\gamma\sigma/(r_1-a\sigma)} \mathrm{e}^{\gamma\sigma/(r_2-a\sigma)} \tag{3.2.3}$$

where $A = 7.049556277$, $B = 0.6022245584$ and $\gamma = 1.2$ give the shape and scale of the potential. The cutoff $a = 1.8$ ensures that all terms in the potential go to

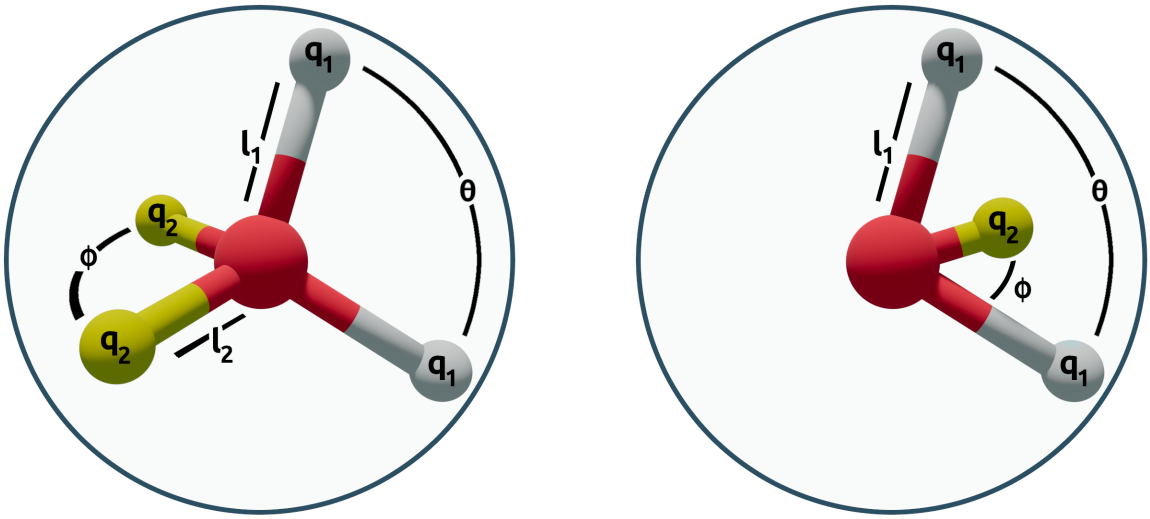


FIGURE 3.2.1 – Schematic representation of the ST2 (left) and TIP4P (right) models, with their geometric parameters and charge location. The blue shell indicates the Lennard-Jones σ parameter.

zero at a distance $a\sigma$. Quadratic cosinus term with $\theta_0 = 109.47°$ favors tetrahedral angles. $\lambda = 23.15$ scales the repulsive three-body interaction term and determines strength of the tetrahedral interaction. $\epsilon = 6.189$ is the strength of pair interaction and $\sigma = 2.3925$ is the particle diameter [3].

The short-range nature of this model render it at least on hundred time faster to compute that all other models evoked here, making it a great tools for quick studies of water properties.

3.2.2 The ST2 model

The ST2 model was one of the first model used to describe water, it was designed to reproduce correctly the radial density function obtained with x-ray scattering experiments [4].

Its geometry consist of a four charge model that agreement the water molecule by two charged lone points, while the oxygen is considered chargeless. The two lone points and the two hydrogen are arranged in a tetrahedral ways, as shown in figure 3.2.1.

The potential energy of the model is the sum of two pair interaction, a Lennard-Jones term and an electrostatic term modulated by a switching function S that goes smoothly from 0 at small distance to 1 at large distance

$$E_{ST2} = \sum_{i,j>i} V_{LJ}(r_{ij}^{OO}) + \sum_{i,j>i} S(r_{ij}) V_e(i,j) \tag{3.2.4}$$

$$V_{LJ}(r) = 4\epsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right) \tag{3.2.5}$$

$$V_e(i,j) = \frac{e^2}{4\pi\epsilon_0} \sum_{a\in\{i\},\, b\in\{j\}} \frac{q_a q_b}{r_{ab}} \tag{3.2.6}$$

$$S(r) = \frac{(r-R_L)^2(3R_U - R_L - 2r)}{(R_U - R_L)^2} H(r-R_L)H(R_U - r) + H(r - R_U) \tag{3.2.7}$$

where ϵ is the strength of the Lennard-Jones interaction and σ is the molecules size. r_{ij}^{OO} are the inter-oxygen distances of the molecules i and j. e is the proton charge and ϵ_0 is the void permitivity. a and b represent the charged particles of the molecules i and j respectively, with q_a, q_b their charge and r_{ab} their distances. $H(x) = 0$ if $x < 0$ and $H(x) = 1$ if $x \geq 0$. All the parameter values are given in table 3.2.1.

ϵ/k_B (K)	σ (Å)	q_1(e)	q_2(e)	l_1 (Å)	l_2 (Å)	θ (°)	ϕ (°)
3.10000	0.31694	0.24357	-0.24357	1.0000	0.80	109.47	109.47

TABLE 3.2.1 – Parameter values of the ST2 models, using same notation as in figure 3.2.1

The ST2 model of water is known to lend an over-structured water, due to its geometry that enforce tetrahedral arrangements. It gives poor prediction about the phase diagram of water, omitting whole phases [71], as shown in figure 3.2.2. Thus results obtained with this potential should be examined with caution.

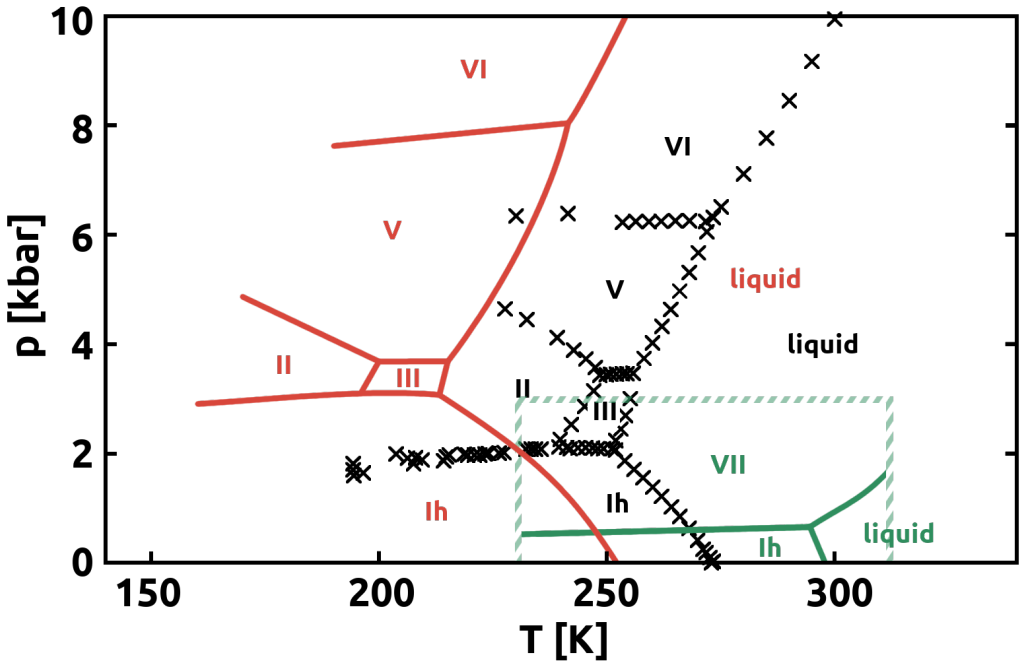


FIGURE 3.2.2 – Comparison of the phase diagram of ST2 (green) and TIP4P/2005 (red) with experiments (black cross), adapted from [5, 71]. The thick lines are phase boundaries that enclose stable phase, their name being indicated by the color corresponding to the model. TIP4P/2005 yield a "correct" phase diagram with all the phases in place, even if largely shifted compared to the experiments. ST2 find that ice VII is the most stable one when going into the high pressure domain, in discrepancy with experiments. Note however that for ST2 only a small part of the (P, T) space was explored, indicated by the green dashed lines.

3.2.3 The TIP4P family

The TIP4P family is a series of models based on the same geometry, but where the potential parameters are fitted to reproduce specific physical properties. Its geo-

metry consists of a three-charges model, where a middle point charge is added in the water molecule plan to effectively represent its small polarization, while the oxygen is chargeless, as shown in figure 3.2.1.

The potential energy of the model has the same functional form as for ST2, except that the electrostatic term is not modulated by a smooth switching function

$$E_{TIP4P} = \sum_{i,j>i} V_{LJ}(r_{ij}^{OO}) + \sum_{i,j>i} V_e(i,j) \tag{3.2.8}$$

where V_e and V_{LJ} are the same as in equations (3.2.5) and (3.2.6). The parameters of the two variants TIP4P/2005 and TIP4P/Ice, employed in this thesis, are given in table 3.2.2. Both models share the same value for the H-O-H angle and the OH distance. Also they both have their negative middle point charge placed along the H-O-H angle bisector so that $2\phi = \theta$ and the value of the charge is $q_2 = -2q_1$.

Model	ϵ/k_B (K)	σ (Å)	q_1(e)	q_2(e)	l_1 (Å)	l_2 (Å)	θ (°)	ϕ (°)
TIP4P/2005	93.2	3.1589	0.5564	-1.1228	0.9572	0.1546	104.52	52.26
TIP4P/Ice	106.1	3.1668	0.5897	-1.1794	0.9572	0.1577	104.52	52.26

TABLE 3.2.2 – Parameters of the TIP4P/2005 and TIP4P/Ice models, using same notation as in figure 3.2.1

TIP4P/2005

The TIP4P/2005 model was developed to reproduce the maximum density of water at $T = 4°$ C and densities of its solid phases. Contrary to most water models, extensive simulations were used to fit its parameters [5].

This model reproduces with a high quality the anomalies of water and the different phases of water [72, 73, 74, 75], even if the phase diagram is distorted compared to the experimental one, see figure 3.2.2. The mains defects are a poor prediction of the dielectric constant, and a diffusion coefficient that is slightly underestimated compared to experiments, even if it has the correct trends [72].

Due to the correct reproduction of the phase diagram and to the high precision in describing the anomalies of water, we will use this model to study the liquid-liquid transition in supercooled water as described in chapter 6.

TIP4P/Ice

The TIP4P/Ice model was developed to reproduce the freezing temperature of water at $T = 0°$ C, and the various densities of its solid phases with the best accuracy, in the same way as TIP4P/2005 [6]. While this model is quite accurate for low-density ices, it struggles to describe accurately very dense ices and their stability domains for pressure higher than 10 kbar [6]. It is not a problem in our case as we will use it to study the homogeneous nucleation of ice, and so we are mainly interested in how the model describes Ice I and its metastable polymorphs – for which TIP4P/Ice is in excellent agreement with experiments – as will be described in chapter 7.

3.3 About the choice of water model

When discussing the mW water model, we mentioned that it is one hundred times faster to compute than more complex models like ST2 or those of the TIP4P family, so one may wonder why we did not pick this model instead of TIP4P/Ice or TIP4P/2005. There are several reasons for this. The first one is that extensive studies have already been conducted on both the liquid-liquid transition [76] and ice nucleation [38] for this model, and it is important to assess if those results are coherent with more realistic models. The second one is that the coarse-grained mW model was designed to reproduce correctly ambient liquid water densities and ice I properties. When computing properties outside of this comfort zone, the model diverges quickly from the properties observed experimentally. For instance, its stable phase diagram is totally off for pressures higher than 1 kbar, producing a new ice phase sc16 instead of the natural ices II, III, V and VI [77]. Furthermore, for all conditions its diffusion coefficient is far higher than in real water [3]. Despite this, its nucleation rate is slower than those found in experiments [78]. So it seems that the mW model is not able to properly describe the kinetic properties of water, even in conditions of pressure and temperature where it gives coherent results for its structural or thermodynamic properties. As one of the goals of this thesis is to generate a set of data from which one could reconstruct the kinetics of water, mW is certainly not appropriate.

4 Describing our systems with collective variables

In order to study phase transformations of matter, the first thing we need is a way to sort the various structural states that matter can reach. For instance, water has more than ten solid phases and it not necessarily trivial to distinguish them. As previously stated, performing classical molecular dynamics give access to the system phase space and so to the positions $x \in \mathbb{R}^{3N}$ of all particles in the system. In theory, from these coordinates we could compute all structural properties of the system. In practice, however, the Cartesian coordinates of all particles are difficult to manage and do not offer directly suitable information to compare two structural states. We need to compute more specific properties from these coordinates, in the form of convenient functions that are called collective variables or structural descriptors, depending if you came from the enhanced sampling or neural network communities. In the following we will mainly use the collective variables terminology to avoid confusion, but the two terms are equivalent.

When we use a collective variable $s(x)$ to describe a transition between two or more states, we speak about order parameter and reaction coordinate : usually, the first term refers to a variable (relatively easy to identify in practice) able to distinguish between the locally stable states at the beginning and at the end of the transformation, while the second term refers to the best possible collective variable, capturing not only the difference between locally stable states but also the detailed transition mechanism (related to the committor, see later).

In this conceptual framework, the free energy (for fixed P, T) will be computed as a function of s, by means of marginal equilibrium probabilities, effectively reducing the dimensionality of the physical problem

$$G(s) = -k_{\mathrm{B}}T \log P(s) = -k_{\mathrm{B}}T \log \left(\int dx P(x) \delta(s - s(x)) \right) \tag{4.0.1}$$

Clearly, such a free energy landscape is much easier to analyse than the potential energy landscape, and it also gives valuable physical insight about the transformation. But as shown in the previous equation, these advantages come at the cost of an integration over phase space, which is not a trivial operation to perform in complex systems like models of materials including thousands of atoms. Both the shape of $G(s)$ and our ability to compute it are strongly dependent of the choice of s. In reality, finding the optimal collective variable is an utterly complex problem, equivalent to gaining a perfect understanding of the transformation process itself, as will be discussed in (4.3).

In principle, one can distinguish two kinds of collective variables, even if the two classes are not always well separated :

— local ones, that describe the states of a single particle. For example, its number of neighbors, if it is solid- or liquid-like, if it is in a cubic or hexagonal ice cell, etc.

— global ones, that describe the system as a whole and generally tell in which states it is. For example if it is solid or liquid, if it is amorphous or ice, etc.

In the rest of this chapter, we will present briefly the various collective variables used in this thesis, starting with the local and then going to the global ones. Each time

we will present in which physical context we used the variables. Among them, the Permutation Invariant Vector (PIV) described in section (4.2.3) is a general approach to define collective variables, which requires very limited or even no knowledge about the system under study. All our enhanced sampling methods described in (5) will use variables based on PIV. The other variables will mainly be used as post-processing analysis tools. We will then present different methods to assess the quality of a reaction coordinate and find optimal ones. Finally we will present how the PIV has been implemented in plumed, a widespread plugin for free-energy calculations and analysis of trajectories, compatible with several MD engines [79].

To compute collective variables, we used plumed or ovito [80] : both are freely available, the first one being open source and community-developed, and can compute a wide range of collective variables, among other features.

4.1 Local collective variables

4.1.1 Coordination number

One of the simplest collective variables that we can compute for a particle, is its number of neighbors in a given range. If one chooses the right physical parameter for the range, this number of neighbors can be thought of as the number of contact between an atom and its surrounding atoms, hence the name coordination number.

If we have an atom i and a group of atoms A, we can compute the coordination number of this i-th atom as :

$$C_i = \sum_{j \in A} \sigma(r_{ij}) \tag{4.1.1}$$

where r_{ij} is the Euclidean distance between the atoms i and j and σ is a switching function that evaluates as 1 if there is a contact and 0 if there are none, so that we effectively count the number of contacts.

We used this collective variable in the study of the liquid-liquid transition of water (see chapter 6). Here all we need to know is that under some specific temperature and pressure condition, water can be in a metastable liquid state. This liquid state is supposed to be composed by two types of liquid : a high density and a low density liquid. We can distinguish them by computing coordination numbers for the oxygen atoms with other oxygens that are within a .34 nm shell. We say that an oxygen with coordination number greater than 4.5 is in high density state and in low density state otherwise [81]. Computation was performed with plumed [79].

4.1.2 The Steinhardt parameters

The Steinhardt parameters are a series of collective variables to measure the degree of ordering of the first shell around an atom (within 0.35 nm). They can be computed for the i-th atom as the norm of a complex vector q_{lm} [82] :

$$Q_l(i) = \sqrt{\sum_{m=-l}^{l} q_{lm}(i)^* q_{lm}(i)} \tag{4.1.2}$$

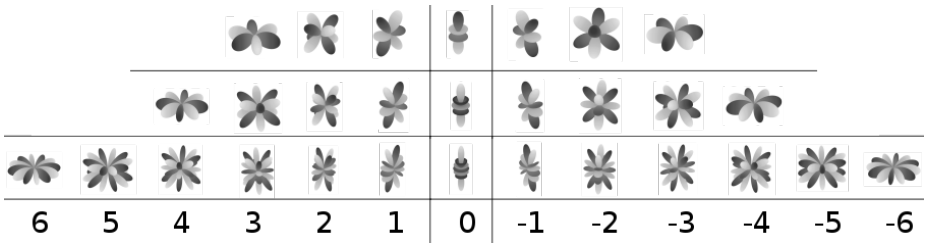


FIGURE 4.1.1 – Visual representation of Y_{3m}, Y_{4m} and Y_{6m}. Light grey area are positive and dark grey area are negative.

where the components of the complex vector q_{lm} are computed as in the following :

$$q_{lm}(i) = \frac{1}{\sigma(r_{ij})} \sum_j \sigma(r_{ij}) Y_{lm}(r_{ij}) \tag{4.1.3}$$

here σ is a switching function that evaluates to 1 if atom j is in the first shell of the atom i and to 0 otherwise, and Y_{lm} is the l-th order spherical harmonic (that is the l-th mode of vibration of a wave in 3 dimensions on a spherical surface). The mathematical definition of spherical harmonics is not particularly enlightening, so instead we provide a visual representation of Y_{6m}, Y_{4m}, and Y_{3m} in Fig. 4.1.1.

In the study of water nucleation only Q_3, Q_4 and Q_6 are used. The Q_6 collective variable is used to tell if oxygen atoms of water molecules are in a crystalline state or not. If Q_6 is less than 0.55 the oxygen is in the liquid state, and it is in a solid state otherwise [83]. It is a 'simple' first approach to monitor the number of solid-like molecules. And by computing the average $\langle Q_6 \rangle$ over all the water particles, one can tell if the whole system is in a solid or liquid state. Computation was performed with plumed [79].

4.1.3 The Chill+ algorithm

If we want more insight about the nucleus structure, we can use the Chill+ algorithm [9]. This algorithm classifies oxygen particles by looking at their local surroundings. In solid phases, water molecules have 4 bonded neighbors. If we take two bonded oxygen particles i and j, that is distant by less than 0.35 nm, we can compute the correlation of the projection of the third-order Steinhardt parameter q_{3m} between them :

$$c(i,j) = \frac{\sum_{m=-3}^{3} q_{3m}(j)^* q_{3m}(i)}{\sqrt{\sum_{m=-3}^{3} q_{3m}(i)^* q_{3m}(i)} \sqrt{\sum_{m=-3}^{3} q_{3m}(j)^* q_{3m}(j)}} \tag{4.1.4}$$

if this correlation is less than -0.8, we say that the bond is staggered, and if it is within the range $[-0.05, -0.2]$, we say that the bond is eclipsed.

Having distinguished this two types of solid bonds, we can identify liquid, interfacial ice, cubic ice and hexagonal ice by simply counting their number of eclipsed or staggered bonds

structure	eclipsed bonds	staggered bonds	neighbors
liquid	N/A	N/A	any
cubic ice	0	4	4
hexagonal ice	1	3	4
interfacial ice	any	2	4

Hence, this algorithm gives us knowledge about both the number of solid-like molecules and their structure. This is of prime importance to study in a detailed way homogeneous nucleation trajectories, as will be discussed in chapter 7. Computation was performed using ovito [80].

4.2 Global collective variable

4.2.1 The largest nucleus size

When we study ice nucleation, it is not enough to have the number of solid like molecules, we also want to know if these molecules form clusters, more commonly named nuclei in the study of nucleation, and the size of the largest nuclei.

Clusters are just sets of connected particles, i.e., if we take two particles, they are in the same cluster if we can follow a continuous path of bonds between them. Once we have identified which water molecules are solid-like with Chill+, we identify clusters with the simple criterion that solid-like molecules less than 0.35 nm apart are bonded. Then we count the number of molecules in every cluster and sort the latter by size to get the largest one.

In fact, once we have the largest cluster, we can compute the largest nucleus size in two different ways. We could only take into account the "strongest" part of the nucleus, not counting the interfacial ice, which is simply the sum of the number of hexagonal N_H or cubic ice molecules N_C in the largest cluster

$$N_{CH} = N_C + N_H \tag{4.2.1}$$

Or we could sum the number of all types of ice, including the number of interfacial ice molecules N_I

$$N_{CHI} = N_C + N_H + N_I \tag{4.2.2}$$

The largest nucleus size will be used to study homogeneous nucleation as will be discussed in chapter 7. Computation was performed using ovito [80].

4.2.2 Cubicity

Once we have identified the largest nucleus, we want a simple measure of its disordering. One simple way to do it is to compute the cubicity C, defined as (keeping the same notation as in the previous section for the number of ice molecules in the largest cluster)

$$C = \frac{N_C}{N_C + N_H} \tag{4.2.3}$$

This number is simply the fraction of molecules that are in cubic ice state, telling us if we are in a purely cubic state ($C = 1$), hexagonal state ($C = 0$), or mixture of the two ($C \in]0, 1[$).

Cubicity will be used to study homogeneous nucleation as will be discussed in chapter 7. Computation was performed using ovito [80].

4.2.3 Permutation Invariant Vector (PIV) and PIV distance

Until now, we have presented collective variables that are specifically designed to solve a given problem. But what if we have little or no knowledge about a system and still want to distinguish the various structures and phases it can reach ? The Permutation Invariant Vector (PIV) is a very general collective variable that aims to solve this question for any atomic system [7].

In general, a collective variable needs to be invariant if we rotate or translate our simulation box, or else it would have no physical meanings, as the states of our system do not change under such symmetries. In addition, it should be invariant under permutation of any pair of identical atoms, as otherwise it would distinguish identical structures that have a different arbitrary labelling of the atoms. For example, in liquid water at ambient pressure and temperature, molecules can freely diffuse and hence permute with each other, even though the states remain the same. Another way of saying this is that we are interested in the topology and not by the topography of our system.

To build the PIV, we start by representing our system by a complete graph weighted by the inter-atomic distances. Then we compute the "adjacency" matrix of this graph, that is a matrix $\mathbf{A}$ with element defined as

$$A_{ij} = \sigma(r_{ij}) \tag{4.2.4}$$

where r_{ij} is the Euclidean distance between atoms i and j and σ is a switching function that goes smoothly from 0 to 1. So contrary to a true adjacency matrix, we have no discontinuity, which will turn out to be important for enhanced sampling methods in chapter 5. This switching function also allows us to set the range of interactions that we consider. Often one will restrict itself to first or second neighbor shells, but if longer range interactions are important they can be easily included.

By construction, this matrix is symmetric and contains all the structural information about the system under study. To build our vector, we first separates it into sub-matrices $\mathbf{B}$ for each pair b of atoms type $\{\alpha, \beta, \ldots\}$ that we are interested in

$$B^b_{ij} = \sigma(r^b_{ij}) \tag{4.2.5}$$

For instance with water, we would have $\alpha = O$, $\beta = H$, with one sub-matrix block for oxygen-oxygen, one for the hydrogen-hydrogen and one for the hydrogen-oxygen interatomic distances, labeled respectively with $b = 1$, 2 and 3. In general if we have N_t type of particles, we will have $N_b = N_t(N_t - 1)/2$ blocks, with N_t of them that contain interatomic distances for same particles types and $N_t((N_t - 1)/2 - 1)$ blocks that will contains interatomic distances for different particles types.

Then we sort the elements inside each block and put them in a vector, that will be invariant under permutation by construction due to the sort operation. Finally we concatenate all these vectors into a single one, to obtain the Permutation Invariant Vector, using the concatenate symbol $\oplus$:

$$\mathbf{v}_b = \text{sort}\left(B^b_{ij}\right) \tag{4.2.6}$$

$$\mathbf{V} = \bigoplus_b \mathbf{v}_b \tag{4.2.7}$$

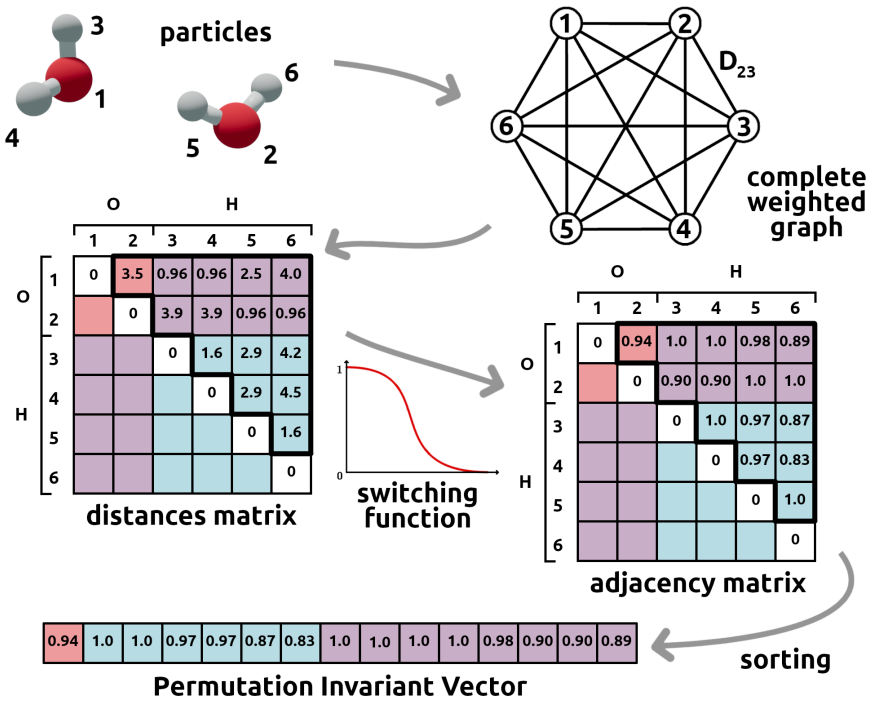


FIGURE 4.2.1 – Construction procedure of the PIV for two molecules of water

These operations are summarized in figure 4.2.1. It is important to note that the sorting of the vector, required for permutation invariance, makes us loose part of the structural information.

If we want to compare configurations with large variations in the volume, we can refine this definition by normalizing the distances with a factor $(V/V_0)^{1/3}$, where V is the box volume and V_0 is a reference, for instance the average volume of the two states we want to compare. We can also include different weights w_b for the different atom pairs, if we think that some pairs are more relevant than others :

$$\mathbf{v}_b = \text{sort}\left(\sigma\left(\left(\frac{V}{V_0}\right)^{1/3} r_{ij}^b\right)\right) \tag{4.2.8}$$

$$\mathbf{V} = \bigoplus_b w_b \, \mathbf{v}_b \tag{4.2.9}$$

The PIV in itself has a high dimension : if we have N particles its size will be $N(N-1)$, so it is not really practical to use it as it is. But from the PIV we can define a very simple metric to distinguish states, that we will call the PIV distance. Let A and B be two structures of interest, and $\mathbf{V}_A$ and $\mathbf{V}_B$ their related PIV. Then the PIV distance is just the squared Euclidian distance between the two PIV

$$D_{AB} = |\mathbf{V}_A - \mathbf{V}_B|^2 \tag{4.2.10}$$

PIV distances will be our main tools in the rest of this thesis to study structural transformations. As discussed in Ref. [84], they are suitable for water and they have the useful property of assuming small values when comparing independent realizations of a same form of water (e.g., two configurations extracted from an equilibrium

trajectory of the liquid), while being large when comparing two configurations belonging to physically different forms (e.g., liquid and crystal or liquid and amorphous). We will use PIV distances to define path collective variables (see next section) within the study of the liquid-liquid transition of water (6). And we will use them both directly and to define path collective variables within the study of homogeneous ice nucleation (7). Computation was performed using plumed [79].

4.2.4 Path collective variables

Path collective variables were specially devised as reaction coordinates for a generic transition between states A and B, provided a suitable metric $\mathcal{D}$ is available to compute distances between any states. The basic idea is to start from a sequence of n reference configurations, that go from A to B with some intermediate states, representing a path that connects A and B [85]. Then, two collective coordinates are defined as

$$S = \frac{\sum_{k=1}^{n} k \, \mathrm{e}^{-\lambda \mathcal{D}(X(t),X_k)}}{\sum_{k=1}^{n} \mathrm{e}^{-\lambda \mathcal{D}(X(t),X_k)}} \tag{4.2.11}$$

$$Z = -\frac{1}{\lambda} \log \left(\sum_{k=1}^{n} \mathrm{e}^{-\lambda \mathcal{D}(X(t),X_k)} \right) \tag{4.2.12}$$

where $X(t)$ represent the current atomic configuration and X_k are the n reference configurations, with $X_1 = A$ and $X_n = B$.

S represents the progress along the reference pathway : when assuming the value 1 it indicates that the current configuration is in state A, while for the value n it is in state B. Z measures the cumulative distance from the reference pathway, large values of Z meaning that the system does not follow the reference path, that can be arbitrary, and follows, typically, a more physical one. So the choice of the reference path is not really crucial, even in combination with biasing potentials, as Z prevents the system from being dragged along unphysical mechanisms. λ is a parameter that controls the shape of the collective variable space, enlarging or shrinking the width of free energy barrier along S when studying a transition process. In practice, a common rules of thumb is to choose it such that $\lambda \mathcal{D}(X_k, X_{k+1}) \approx 2.3$, with $\mathcal{D}(X_1, X_2) \simeq \cdots \simeq \mathcal{D}(X_{n-1}, X_n)$.

In this thesis, following Ref. [8], we mainly used path collective variable with PIV distance D as the metric, and only with the initial and final states of the transition we want to study as references ($n = 2$), in order to avoid making any guess about the path of the transformation. In this scheme, S and Z take the simpler form

$$S = \frac{1 \, \mathrm{e}^{-\lambda D_{X(t)A}} + 2 \, \mathrm{e}^{-\lambda D_{X(t)B}}}{\mathrm{e}^{-\lambda D_{X(t)A}} + \mathrm{e}^{-\lambda D_{X(t)B}}} \tag{4.2.13}$$

$$Z = -\frac{1}{\lambda} \log \left(\mathrm{e}^{-\lambda D_{X(t)A}} + \mathrm{e}^{-\lambda D_{X(t)B}} \right) \tag{4.2.14}$$

where are A and B will be a low density liquid and high density liquid state during the study of the liquid-liquid transition (chapter 6), and will be a liquid and purely hexagonal ice state during the study of homogeneous ice nucleation (chapter 7). Computation was performed with plumed [79].

4.2.5 Commitment probability

The commitment probability, more commonly named committor, is a special kind of reactive coordinate able to measure the progress of a transformation. Consider a system of N particles, that can reach two states A and B, being in the configuration $x \in \mathbb{R}^{3N}$. The committor $\phi_B(x)$ is defined as the probability of the system to reach first B instead of A for a set of many trajectories initiated at x with an equilibrium distribution of initial velocities. $\phi_B(x)$ varies smoothly between 0 and 1 [10, 11]. Thus $\phi_B(x) = 0$ means that the system is in state A and so it will never reach B first. On the contrary $\phi_B(x) = 1$ means that the system is in state B and so it will ever reach B first. Symmetrically, we can define $\phi_A = 1 - \phi_B$. When $\phi_a(x) \approx \phi_B(x) \approx 0.5$, x is a transition state [86].

Here is an important point that we want to stress : in condensed matter we often study systems where the two states will be separated by high free energy barriers. The higher the barrier, the lower the probability per unit time of observing a spontaneous transition where the system crosses the barrier. This effectively means that except close to the top of the barrier, the committor will evaluate to values very close to 0 or 1.

What makes the committor so special is that it contains rich information about the kinetics of the transformation, telling us directly what is the likely fate of a configuration. But contrary to all the other variables defined until here, the committor cannot be computed directly as an explicit mathematical function of the atomic coordinates. One way to estimate it for a configuration is to propagate several molecular dynamics trajectories with different velocities, and count how many of them reach B before A, to effectively sample the probability [86]. Obviously, it would be utterly expensive to estimates precisely the committor for all possible atomic configurations x in a system formed by more than a few particles. This is due to two reasons : on one side, the immense number of possible configurations in a system of hundreds of atoms, and on the other side the large number of trajectories that are needed to estimate $\phi_B(x)$ far from the transition state region. For instance, with $\phi_B(x) \sim 10^{-6}$ we would need to generate more than 10^6 trajectories. As a result, the committor remains an important conceptual tool, while it can be estimated in practice only for a limited number of configurations close to the transition state, by generating few dozens of trajectories for each trial configuration.

4.3 Quality of reaction coordinates

As said, an important use of collective variables is to describe transition processes. Given a reaction coordinate s that is an explicit function of configurations $x \in \mathbb{R}^{3N}$ of our system, we can project the free energy onto it. In practice, one will estimate the equilibrium distribution $P(s)$, either from brute force molecular dynamics or using some enhanced sampling methods, and from there compute the free energy

$$G(s) = -k_{\mathrm{B}} T \log(P(s)) \tag{4.3.1}$$

The question is, can we assess that this projection will preserve the important features of the transition (correct number of local minima and barriers, as well as their relative elevation) ? And can we find an optimal reaction coordinate that describes "perfectly" the transition ?

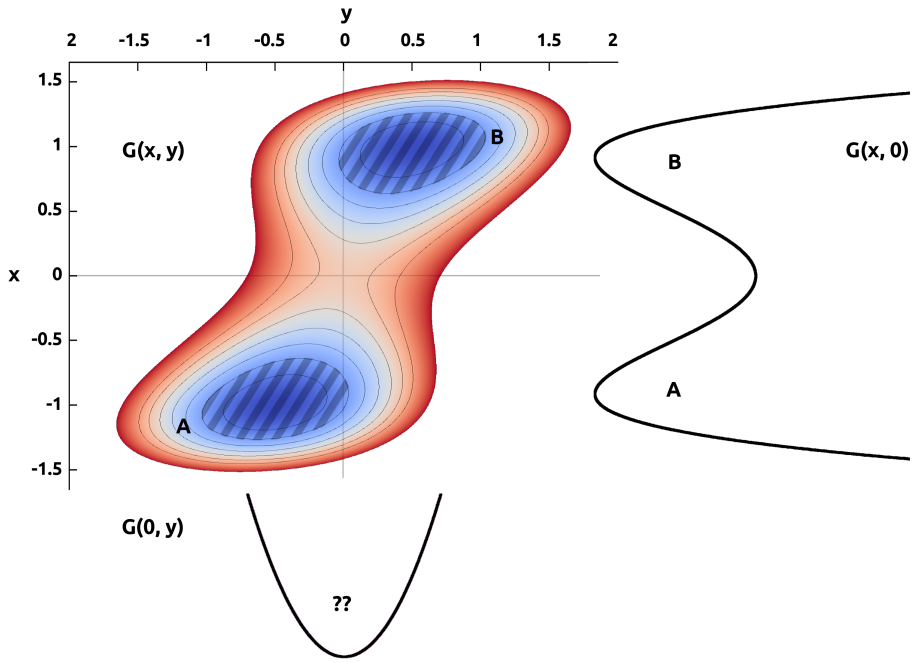


FIGURE 4.3.1 – A fictitious 2D free energy landscape and its 1D projections. Here y is certainly a bad reaction coordinates as it is not able to distinguish states A and B. On the contrary x seems to grasp the important features of G, and may be a good candidate.

If we study a transition between states A and B, a basic requirement is that the reaction coordinate must be able to distinguish A and B. For instance $\langle Q_6 \rangle$ cannot distinguish low-density liquid and high-density liquid [23], but is very competent to distinguish liquid and ice [83]. Thus it is clear that it cannot be used to study the liquid-liquid transition, but it is not clear if it can be used to describe properly the nucleation of water in all its relevant details.

It is always possible to project the free energy on any coordinate, but in unfavorable cases the resulting landscape can suggest misleading physical insight [86]. In figure 4.3.1, we have illustrated how this projection could result in a misleading free energy landscape. It is important to note, moreover, that in real world cases we are reducing the dimensionality way more dramatically, as we often go from a $3N$ dimensional space with N of the order of hundreds or more, to a 2 or 1 dimensional space.

This is of prime importance when we consider enhanced sampling methods, as despite their differences, they all are likely to be effective if a good reaction coordinate is known. Conversely, a poor definition of reaction coordinate can seriously hamper all of those algorithms, leading to non-optimal transition mechanisms and preventing free energy landscapes to converge [46].

The main idea employed to assess the quality of a reaction coordinates s is to analyze the projection of the committor onto it : $\phi_B(s)$. In fact, ϕ_B is widely considered the optimal reaction coordinate for every reaction, hence if s is a good reaction coordinate it should map properly to ϕ_B and in its space the transition state should display a committor distribution peaked around 0.5. Otherwise this means that s is a poor estimate of the ideal reaction coordinate [15, 11, 87, 88, 89]. As

already mentioned, this means that, usually, the quality of a coordinate can only be assessed a posteriori, once we have collected enough data about the transition under study to estimate its committor.

4.3.1 Maximum likelihood optimization

Besides the validation of reaction coordinates, the committor can also be used to optimize reaction coordinates. The maximum likelihood optimization approach was developed with this aim [12, 13, 88, 89]. In this scheme, a large ensemble of points on top of the barrier is sampled through the use of aimless shooting algorithm (see section 5.2), which works independently of the choice of reaction coordinates (as long as they distinguish correctly the metastable states A and B). From this ensemble of shooting points, we only keep information about in which state the endpoints of the generated trajectory are, and so if it is a reactive one. For instance, a trajectory that starts in A and ends in B is reactive, but one that starts and ends in B is not.

After harvesting all this information, one can specify a set of collective variables $s = s_1, \ldots, s_m$ to be tested. There are no requirements on the s_i, so they can have different units and scale. All the candidate variables are evaluated at each previously collected shooting point. As both the fates of the trajectories and the value of s_i at the shooting points are known, we can evaluate the transition path distribution projected on each variable and compute its likelihood with a model based on the committor

$$L(s_i) = \prod_{k=1}^{N_r} \phi_B(s_i(x_k)) \prod_{k=1}^{N_{nr}} (1 - \phi_B(s_i(x_k))) \tag{4.3.2}$$

where N_r is the number of reactive trajectories, N_{nr} is the number of non-reactive trajectories and x_k is the configuration of the k-th shooting point. This likelihood effectively tells us how well the committor projected on the candidate variable fits the ideal committor distribution. By taking $l(s_i) = -\ln L(s_i)$ and sorting it in increasing order, the lowest value will give us the best reaction coordinate among the candidates one. In this thesis we only used this scheme to measure the quality of our variables, but the approach was also devised to optimize the likelihood of a variable constructed as a flexible combination of the trial variables s_i, hence the name.

Finally, one should appreciate that finding the optimal reaction coordinate is more than a technical issue : it often leads to a deeper understanding of the nature and driving forces of a transformation process, hence it is desirable even when enhanced sampling is not needed.

4.4 Development of PIV in plumed

The first version of the code to compute the PIV with plumed was written by Silvio Pipolo, and it was designed to take several reference structures as input and to compute the distance of the current state from these references as output. It successfully accomplished this task, allowing its use in several cases [8, 90]. But it suffered from a tedious user interface, linked to how plumed manages its collective variables with several outputs. In practice this means that the input file size one needs to provide grew linearly with the number of references and was affected by bugs beyond 3.

During this thesis the code was upgraded with different objectives : the first one was to enhance performances, second one was to allow and ease the use of more than 2 reference structures, the third one was to simplify the format of the input to diminish the risk of user mistakes, and last one was to clean-up the code to facilitate further development in the future. Here we will present how the code works, broadly, and how it was upgraded, assuming little familiarity with `c++`. We will mainly present its most tricky parts, which are how we sort the PIV and how we compute its derivatives.

4.4.1 User interface

The first version of PIV implementation used several tricks to circumvent use of collective variables with several components in plumed. The main idea is that if the user wants N references, the code will need to create N PIV objects with different structure references but who share common data through the use of `static` keyword. This means that the user needs to repeat N times the same PIV command in its plumed.dat input file, as shown in table 4.4.1. It also implies that the code *could* misbehaved in unexpected and uncontrolled way when used with several parallel threads, as static variables may be updated concurrently and thus are not thread safe.

Upgrading the interface implied a deep modification of the code to handle multiple components with only one object. The current implementation counts how many references are given in the input and adds a plumed component for each one by calling `void addComponentWithDerivatives (const std::string&)`, instead of using a single call to `void addValueWithDerivatives ()`. This simple change implies several modification of the code, as all the internal variables used to store data about structure references need to become arrays. Notably, it implies to change how derivatives are computed, having an array of `double` for each structure references instead of one by PIV object. Globally it renders the code more complex, even if more robust and with less sources of error as we get rid of all `static` variables, but externally this simplifies a lot the input format, as shown in table 4.4.2.

In fact if you compare the two inputs, you may see that there are also several keywords that disappear. For `PIVATOMS` this is because we can compute it directly from the `ATOMTYPES` entry. For `SORT` now by default all block are sorted, so it only needs to be specified if you do not want to sort any block (beware that we loose the permutation invariance in this case). `PRECISION` has now a default value of 1000, which is a good compromise between precision and performance for most practical cases. The `NLIST` keyword was removed as the current implementation only works with neighbor lists. For the `VOLUME` keyword, we made the choice to use the average volume of the reference structures instead of one provided by the user, as it should be appropriate in all cases. For `NL_CUTOFF`, `NL_STRIDE` and `NL_SKIN` we made the choice to use a unique value for all blocks instead of one for each, as we did not meet a practical case where the various blocks required different length scales and behaviors for the neighbor lists. Also it simplify the input and the code, reducing potential sources of errors.

```
PIV ...
  LABEL=d1
  REF_FILE=liquid.pdb
  ATOMTYPES=OW1,HW
  PIVATOMS=2
```

```
  SFACTOR=1.0,0.2,0.2
  SWITCH1={RATIONAL R_0=0.7 MM=12 NN=4}
  SWITCH2={RATIONAL R_0=0.7 MM=12 NN=4}
  SWITCH3={RATIONAL R_0=0.7 MM=12 NN=4}
  SORT=1,1,1
  PRECISION=1000
  NLIST
  VOLUME=24.34874
  NL_CUTOFF=1.2,1.2,1.2
  NL_STRIDE=10,10,10
  NL_SKIN=0.1,0.1,0.1
... PIV
PIV ...
  LABEL=d2
  REF_FILE=ice.pdb
  ATOMTYPES=OW1,HW
  PIVATOMS=2
  SFACTOR=1.0,0.2,0.2
  SWITCH1={RATIONAL R_0=0.7 MM=12 NN=4}
  SWITCH2={RATIONAL R_0=0.7 MM=12 NN=4}
  SWITCH3={RATIONAL R_0=0.7 MM=12 NN=4}
  SORT=1,1,1
  PRECISION=1000
  NLIST
  VOLUME=24.34874
  NL_CUTOFF=1.2,1.2,1.2
  NL_STRIDE=10,10,10
  NL_SKIN=0.1,0.1,0.1
... PIV
PRINT ARG=d1,d2 STRIDE=1 FILE=cv_piv.dat FMT=%15.6f
```

TABLE 4.4.1 – Example of a minimal .dat file required to compute the PIV distance from two references at every time step with the old interface.

```
PIV ...
  LABEL=piv
  REF_FILE1=liquid.pdb
  REF_FILE2=Ih.pdb
  VOLUME
  ATOMTYPES=OW1,HW
  SFACTOR=1.0,0.2,0.2
  SWITCH1={RATIONAL R_0=0.7 MM=12 NN=4}
  SWITCH2={RATIONAL R_0=0.7 MM=12 NN=4}
  SWITCH3={RATIONAL R_0=0.7 MM=12 NN=4}
  NL_CUTOFF=1.2
  NL_STRIDE=10
  NL_SKIN=0.1
... PIV
PRINT ARG=piv.d1,piv.d2 STRIDE=1 FILE=cv_piv.dat FMT=%15.6f
```

TABLE 4.4.2 – Example of a minimal .dat file required to compute the PIV distance from two references at every time step with the new interface.

4.4.2 Counting sort algorithm

The construction of the PIV is done by blocks of pairs of atom types, each of these blocks being computed separately. The construction of a block and its sorting is parallelized with MPI, using the counting sort algorithm. This works as follow : each thread computes an histogram with `NPRECISION` bins of the atom pair distances normalized by the user-defined switching function. This histogram will effectively count how many times a specific normalized distance occurs, being by construction sorted. To reconstruct the full PIV from it, we gather all the histograms computed in parallel into one and we expand it, omitting the first bin that contains numerous zeroes to gain computation time. Table 4.4.3 shows a simplified serial version of the algorithm as implemented in the code.

```
// cache each PIV block separately
auto currentPIV = std::vector <std::vector <double>> (mBlockCount);
for (unsigned bloc = 0; bloc < mBlockCount; bloc++) {
  // count occupancies
  auto orderVec = std::vector<int> (mPrecision, 0);
  for (unsigned atm = 0; atm < mBlockAtoms[bloc]->size(); atm += 1)
   {
    // compute pair distance
    auto atomPair = mBlockAtoms[bloc]->getClosePairAtomNumber (atm);
    auto position0 = atomPosition (atomPair.first.index ());
    auto position1 = atomPosition (atomPair.second.index ());
    auto pairDist = distanceAB (position0, position1);
    // transform distance with Switching function and then into int
    auto df = double (0.);
    auto vecInt = static_cast<int> (
      mSwitchFunc[bloc].calculate (pairDist.modulo() * mVolumeFactor
   , df)
      * static_cast<double> (mPrecision - 1) + 0.5
    );
    // keep distance count
    orderVec[vecInt] += 1;
  }
  // reconstruct the full PIV
  for (unsigned i = 1; i < mPrecision; i++) {
    for (unsigned m = 0; m < orderVec[i]; m++) {
      currentPIV[bloc].push_back ( double(i) / double(mPrecision -
   1) );
    }
  }
}
```

TABLE 4.4.3 – Simplified serial version of the construction of PIV using counting sort algorithm. `mBlockAtoms` type is `std::vector< std::unique_ptr <NeighborList>>`, it contains the list of atoms for each PIV block. `NeighborList` is a part of the plumed core, and contains atoms positions. `mSwitchFunc` type is `std::vector<SwitchingFunction >`, with `SwitchingFunction` a part of the plumed core, it contains the user defined switching function.

4.4.3 Computing the derivatives

To allow use of PIV with enhanced sampling methods that add biases, we need to compute its derivatives. In practice if $\mathbf{V}$ is the current structure PIV and $\mathbf{V}^r$ is the reference structure PIV, the distance is computed as

$$D = \sum_{b,a}^{N_b,M_b} w_b(V_{ba} - V^r_{ba})^2 \tag{4.4.1}$$

where N_b is the number of blocks, M_b the size of the block b and w_b the weight of the block b. Its derivative with respect to the i-th atomic position $\mathbf{x}_i$ is

$$\partial_{\mathbf{x}_i} D = \sum_{b,a} 2w_b(V_{ba} - V^r_{ba})\partial_{\mathbf{x}_i} V_{ba} \tag{4.4.2}$$

and

$$\partial_{\mathbf{x}_i} V_{ba} = \partial_{\mathbf{x}_i} \left[\text{sort} \left(\sigma \left(\left(\frac{V}{V_0} \right)^{1/3} r^b \right) \right)_a \right] \tag{4.4.3}$$

$$= \partial_{\mathbf{x}_i} \left(\left(\frac{V}{V_0} \right)^{1/3} \sqrt{(\mathbf{x}^b_k - \mathbf{x}^b_l) \cdot (\mathbf{x}^b_k - \mathbf{x}^b_l)} \right) \partial_r \sigma(r) \tag{4.4.4}$$

$$= \left(\frac{V}{V_0} \right)^{1/3} (\mathbf{x}^b_k - \mathbf{x}^b_l) \left(\delta_{ik} - \delta_{il} \right) \partial_r \sigma(r) \tag{4.4.5}$$

where k and l are the indices of the a-th element of the PIV block V_b after sorting. $\partial_r \sigma(r)$ is the derivative of the switching function and depend of it specific form (in practice it is managed by plumed itself, so we do not really care). δ_{ij} is the Kronecker's symbol, that evaluates to 1 if $i = j$ and to 0 otherwise.

As it can be seen, this implementation implies a lot of bookkeeping, because we need to track down the indices of all atoms when we compute the PIV. To do this, we construct with the same counting sort algorithm two supplementary vectors `atmI0` and `atmI1`. They are defined such that if we have inserted the normalized distance r_{ij} between atoms i and j into the PIV, i will be stored in `atmI0` and j in `atmI1`. On this part of the code, a series of small optimizations were made, such that in the end the overall code is at least 4 time faster.

5 Enhanced sampling methods

As mentioned previously, pure molecular dynamics gives access to a wealth of microscopic information. Apart from the accuracy of the force fields used to describe the given material, its main limitation comes from the typical time scale that one can reach with nowadays computers. To evaluate equilibrium properties of a specific state this is often not a problem. But when we want to explore the available metastable states or study their relative stability and intereconversion kinetics, this is not enough.

To describe the states stability of a system in the NPT ensemble, we use the Gibbs free energy G. G is an equilibrium quantity, in which each minimum represents a metastable state (the global minimum being the stable one) and minima are separated by free energy barriers, which effectively trap the system in each state. With a small probability, the system can spontaneously cross the barrier over a short period of time and fall in another metastable state. This crossing probability is exponentially decreasing with the height of the barrier. Figure 5.0.1 presents roughly the typical time scale reachable by molecular dynamics, and the typical time scale needed to see the occurrence of a rare transition event. As you can see, to study transitions in material, basic molecular dynamics is of relative use.

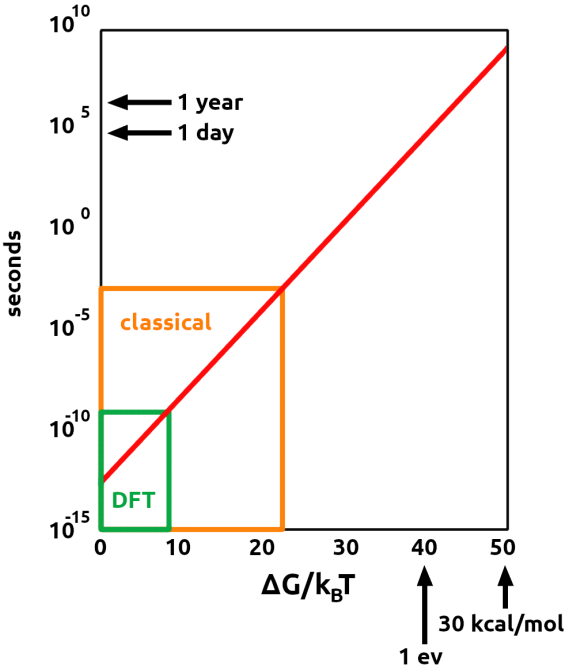


FIGURE 5.0.1 – Theoretical time scale needed to overcome a fixed free energy barrier. Typical time scale reachable by quantum based (DFT) and classical molecular dynamics simulation.

To overcome this major problem many methods were and are still developed to enhance the sampling of the transition area. There are three distinct features that can be achieved by enhanced sampling methods, in increasing order of complexity

— exploration of the relevant structures on the free energy landscape, to obtain the available metastable states

— quantitative reconstruction of the free energy landscape, to get the equili-

brium probability of metastable states
— precise reconstruction of the characteristic transition times between metastable states of the system, i.e. the kinetic properties.

For the first two features there are a wide arrays of methods available, that may yield quantitative and qualitative results if used properly. The ones we used are presented in (5.1). For the last one however there are far fewer methods developed and tested, due to its higher complexity. The methods we used are presented in (5.2). The enhanced sampling methods that we adopted can be separated in two classes : adding bias cleverly along the reaction path or shooting a large number of trajectories from the top of the barrier.

Adding biases

For this class of method, we reduce the dimensionality of G by projecting it onto some collective variable $s(x)$, $x \in \mathbb{R}^{3N}$, that is supposed to be a good reaction coordinates, as described in (4.3). Now by adding a bias $V_B(s,t)$ to $G(s)$, we can change in a favorable way the dynamics of the system. An edge case is if we have $\forall s, V_B(s,t) + G(s) \approx 0$, in which case the transition are not hampered by any barrier and the system can freely diffuse from one state to another.

In practice we do not know $G(s)$ in advance, so the aim is to build $V_B(s,t)$ in such a way that it will either reconstruct directly the free energy after some time, or it will allow to accumulate enough statistics to reconstruct a proper estimate of $P(s)$. Metadynamic is of the first type and umbrella sampling of the second.

It's important to note that for this types of methods s need to be a good reaction coordinates and to be continuous. As to compute the bias forces exerted on every particles we will need to derive the bias potential. For the particles i we will have

$$F_i = -\partial_{x_i} s(x) \partial_s V_B(s) \tag{5.0.1}$$

This is why in the definition of PIV, we use a smooth continuous switching function and not a sharp discontinuous one.

Shooting from the top

For this class of methods, we will propagate (or shoot) several molecular dynamics simulation starting from a specific configurations, typically on the top of the free energy barrier that separates the two states (hence the name). Then we will couple these shootings with some algorithm that only keep the relevant propagated trajectories, in such a way that we can sample precisely the free energy barrier and its kinetic.

All these methods rely on the ability to tell in which metastable state we are and if we are in a transition state, through the use of a collective variable. If the ability to distinguish two state is not a specifically harsh requirement, the ability to describes properly transition states is one, as discussed in (4.3). In fact for transition path sampling the efficiency of the algorithms is directly related to the choice of variable.

5.1 Free energy exploration and reconstruction

In this thesis we used two methods to explore or reconstruct the free energy landscape. Umbrella sampling was used for precise reconstruction of the free energy. Metadynamic was used for exploration and rough estimation of the free energy. Here we will present the general theory, as the contextual utilization settings will be detailed in chapter 6 or 7.

5.1.1 Umbrella sampling

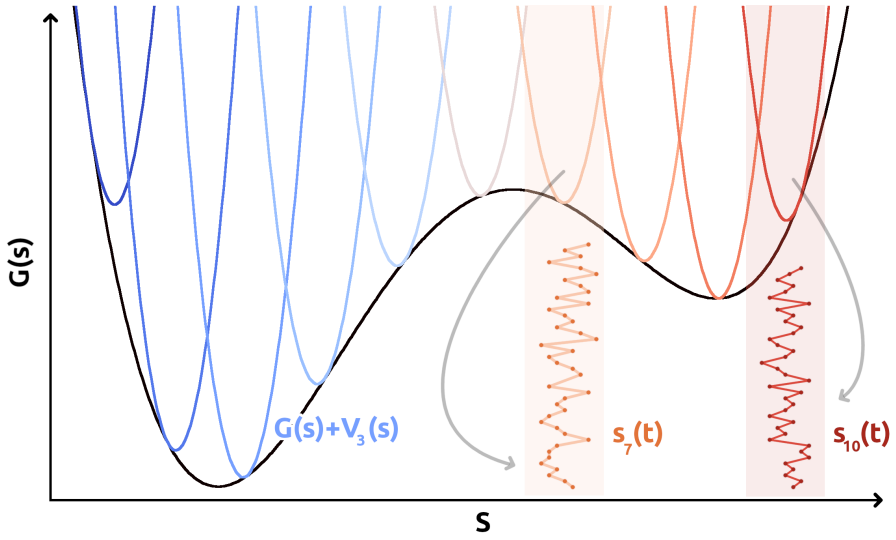


FIGURE 5.1.1 – Schematic representation of the umbrella sampling method, where we add strong quadratic potential (umbrella) to restrain trajectories into small windows sampled separately. In real use the quadratic bias will be sharper and more numerous.

Umbrella sampling is a method that allow one to finely reconstruct the free energy landscape. The principle is to launch several simulation for which we added a constant quadratic bias potential, often called umbrella due to their shape, along a reactive path defined by a collective variable s [14]. The i-th bias is defined as

$$V_i(s) = \frac{k}{2}(s - s_i)^2 \tag{5.1.1}$$

If we tune k to have strong enough bias, the trajectory will be forced to sample precisely a small region centered around s_i that we will call a window. In the i-th window we will sample a distribution

$$P_i(s) = \frac{P(s)\,\mathrm{e}^{-V_i(s)/k_\mathrm{B}T}}{\int \mathrm{d}s' P(s')\,\mathrm{e}^{-V_i(s')/k_\mathrm{B}T}} \tag{5.1.2}$$

We can then combine all the sampled distribution P_i to reconstruct an estimate of the underlying total distribution P, and thus of G. The general principle of the umbrella sampling method is illustrated in figure 5.1.1.

To reconstruct the total distribution P from the sampled distribution P_i we use the weighted histogram analysis method (WHAM) and its bayesian version the

multistate Bennet acceptance ratio estimator (MBAR) [91, 92]. The idea of this two methods is to build an estimator of the free energy, based on the sampled P_i, that we will solve in an self-consistent way. In practice we have an equation $G_i = f(G_i)$, where f is some function, that we solve by iteration using the last set $\{G_i^n\}$ to produce a new estimated set $\{G_i^{n+1}\}$. The difference between the two resides in the use of a discretized, using histograms, or continuous density of state. In the limit that histograms bin width go to zero, the WHAM method became equivalent to the MBAR method [92].

The fact that each windows are sampled separately make this method intrinsically parallel, a major advantage on current architectures of supercomputers. It also means that its easy to add statistics to a specific windows if its convergence is not optimal. Combined with its rigorous mathematical foundation, this imply that umbrella sampling is a very powerful general technique to estimate the free energy. Its main limitations consist in the necessity to have an initial continuous pathway to initialize the various windows, and in the necessity to carefully monitor the dynamics in the different windows, to avoid the exploration of disconnected regions in configuration space [93]. In practice this method will be used in tandem with another enhance sampling technique able to generate such trajectories, like seeding or metadynamics.

Also umbrella sampling is a double-edged blade that should be used with some special caution, as its convergence is not trivial to assess. Contrary to other methods like metadynamics where convergence issues appear more clearly, the use of WHAM or MBAR to estimate free energy from umbrella sampling simulations will always deceivingly converge from a numerical viewpoint, as long as the collective variables distributions overlap, even if the underlying simulations are totally un-physical.

There are three main sources of failure for umbrella sampling : k in equation 5.1.1 is too small ; each windows are not sampled for long enough ; the continuous pathway used to launch umbrella sampling is shitty. The first one is easy to verify, as we just need to check overlap of the distributions P_i. The second one can be verified by computing auto-correlation time of the collective variable s in every windows, defined as

$$\tau_{s_i} = \int_{t_0}^{t_{max}} \mathrm{d}t \frac{\langle \delta s_i(0) \delta s_i(t) \rangle}{\langle \delta s_i^2 \rangle} \qquad \text{with} \qquad \delta s(t) = s(t) - \langle s \rangle \tag{5.1.3}$$

for $t_{max} \gg \tau_{s_i}$ (typically 1 or 2 order of magnitude), we can "safely" say that the umbrella are converged. For the last one, it is often linked to a bad choice of reactive coordinates. How to assess the quality of the coordinate chosen is discussed in section (4.3).

5.1.2 Metadynamics

Metadynamic is a method devised to both reconstruct the free energy landscapes and explore it. Contrary to umbrella sampling where we added constant biases, here we will generate dynamically a time-dependent bias, which will be specific to the simulation. Let say we have a collective variable s and a deposit time step τ. Every τ times step, we we will add a small potential of gaussian shape to our system, at its current location $s(t = n\tau)$. In this way we will progressively construct a potential V that will have the opposite shape of the free energy G, allowing both to overcome large free energy barrier and to reconstruct G [39]. After n gaussian deposit, the

potential will be

$$V(s,t) = w \sum_{k=1}^{n} \exp\left(-\frac{(s - s(k\tau))^2}{2(\sigma_s)^2}\right) \tag{5.1.4}$$

where w and σ_s are the height and width of gaussian potential. The general principle of this method is illustrated in figure 5.1.2.

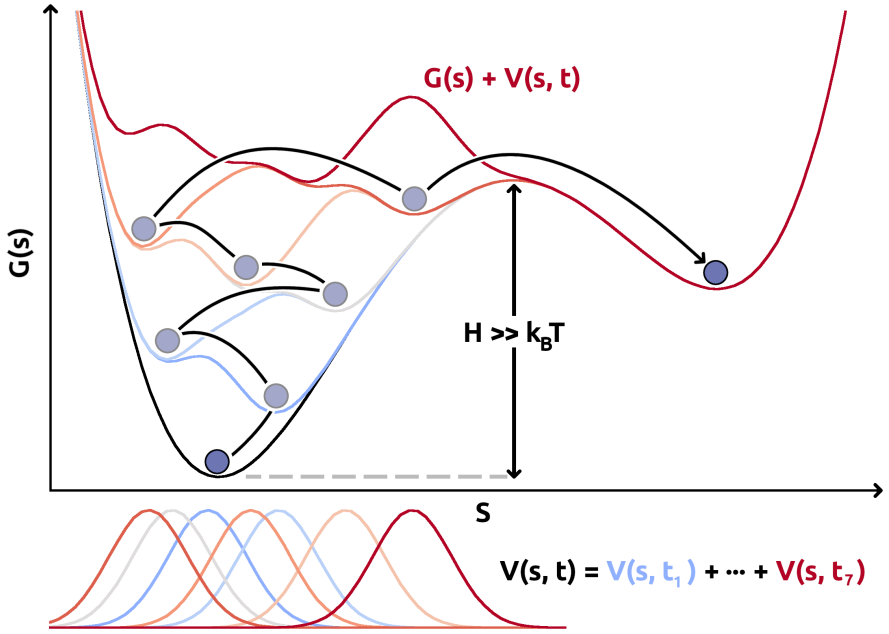


FIGURE 5.1.2 – Schematic representation of the metadynamic method, where we add small gaussian potential at the current system position to fill-up free energy minima and allow the system to overcome large free energy barrier of height H. Here gaussian are very large and high to ease the illustration.

When we always add the same constant gaussiann potential, we speak about ordinary metadynamic. In this case, if the chosen collective variables are able to distinguish the transition state as discussed in (4.3), the time dependent potential will converge toward the inverse of the free energy within numerical uncertainty when $t \to \infty$. But if the collective variable is not, it will lead to a biased free energy estimation [94]. To circumvent this issue, one can use well-tempered metadynamic, a variant in which the gaussian deposit size is decreased in a specific way over the course of the simulation [95]. In this case, the time dependent potential will converge toward the inverse of the free energy, regardless of the capacity of the collective variable to describe the transition state [96]. However its important to note that in cases where the collective variables are particularly badly chosen, ordinary metadynamics may still explore a large part of the configuration space, whereas well-tempered metadynamics may be trapped in a delusive free energy minima that is artificially created by the dimensionnal reduction.

As you may guess, in practice convergence may not be reachable. This is in part due to the limited time scale available by molecular dynamics simulation. But it is also due to the overgrowing complexity to compute forces with such time-dependent potential, as every τ time step we add a new potential and its set of derivatives. Also

one need to not add too frequently the gaussian biases, in a way that the system can relax between two deposits. In an extreme case, if we add the bias with the same time step as the one of the simulation, the system will not be able to move at all and we will just pill-up gaussian on top of each other, breaking the simulation due to the added energy.

Linked to this convergence problem, it may be difficult to determine when to end a metadynamics run. It is especially true for ordinary metadynamic, as in the long-time limit the recovered free energy surface fluctuates around the actual free energy, and the magnitude of the fluctuations is controlled by the rate at which the small gaussian functions are added to the potential energy [94, 97, 98]. If the system has a slow diffusion over the collective variable, this can typically lead to artificial hysteresis, where the system oscillate between its metastable states as if they were a barrier, even though it has been exceeded long ago. Using well-tempered metadynamics may solve this issue, as the height of the gaussian is rescaled every time step by a bias factor, ensuring more smooth and guaranteed convergence in a finite time [95]. In both cases, the main way to control the convergence and to compute error of the free energy estimation, is to use block averaging techniques. The idea is to split the simulation into blocks of the same size, then by looking at the variation of the average bias potential in each blocks, one can estimate the error [96].

For ordinary metadynamic the three parameter of the simulations (deposit time, width and height of the gaussians) should be chosen with care to have smooth exploration or convergence within the time limit of the simulation. The choice of the width is an easy one, as we generally have information about the available range of the reaction coordinate used to study the transition. For the height and the deposit time, it is a trade-off between precision and speed of exploration, as high gaussian will allow one to explore the configuration space quickly, but with poor estimate of the underlying free energy landscape. For well-tempered metadynamic the reduction of the gaussian height is further controlled by a bias factor, which should be chosen such that the system can cross the free energy barrier in the time scale of the simulation.

Since its invention, ordinary metadynamic and its variants have been successfully applied to study many problems. It can be used to directly compute free energy profile, or as a purely explorative method. In the latter case, it can be either to explore the available configurations of a system and find its metastable states, or to generate initial reactive trajectories, a non-trivial task in many problems in complex system, before resorting to other enhance samplings methods like umbrella sampling or transition path sampling [99].

Following the second path, in this thesis we will mainly use ordinary metadynamic to generate reactive trajectories before use of a more predictable and parallel method, umbrella sampling, so we will not worry too much about its convergence. In all of the rest of this thesis we will simply refer to the ordinary metadynamic as metadynamic.

5.2 Transition path sampling

In the previous section we presented a series of methods that allow one to overcome large free energy barrier and to reconstruct the whole free energy landscape of any transformation, given that we have the right collective variables to describe

it. Free energy give us a lot of information about stability or metastability, but it doesn't inform us about the dynamic properties of our system, as the bias introduced to sample it spoils all kinetic information. And these are most valuable, like the transition rate at which our system transform from one state to another.

So to get information about the kinetic, we need to resort to unbiased methods that are still able to sample transition. In the last decades, a whole zoology of such methods were developed by the scientific community, with their drawback and advantages. Among them we mainly used transition path sampling methods.

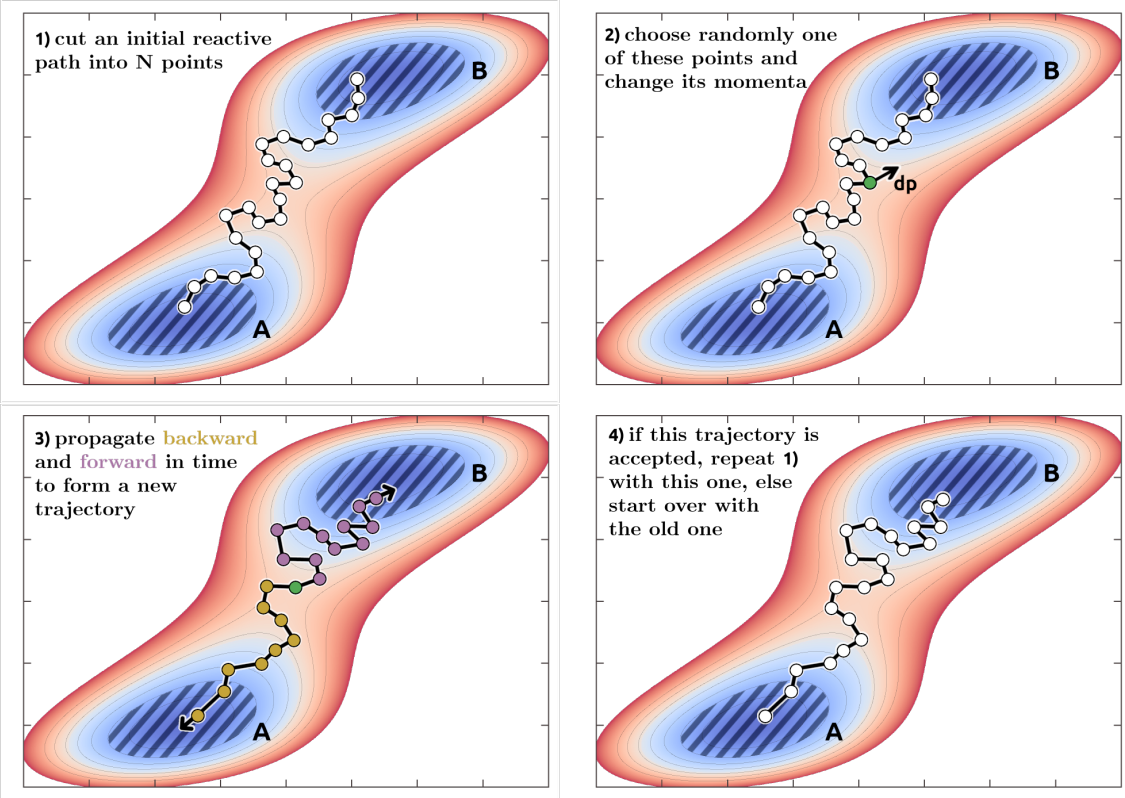


FIGURE 5.2.1 – Schematic illustration of the transition path shooting general algorithm.

Before entering into the details, just a bit of terminology. *Reactive trajectory* or *transition path* is used when molecular dynamics trajectory connect two different states. For instance when we study crystallisation of water, a reactive trajectory/-transition path is one that goes from liquid to ice, or the reverse. *Propagating the system* just means running a molecular dynamics simulation with fixed time length. *Shooting point* are the initial structure from which trajectory will be propagated, i.e. a specific set of cartesian coordinates of our system. *Markov processes* are random models that describe a sequence of possible events and for which the probability of an event is only dependent of the previous event state.

Transition path sampling is a series of methods that tries to overcome disparity in time scales through importance sampling in trajectory space. This means that the objectives of these methods is to sample distribution of transition path through the use of short trajectories. Instead of running infinitely long molecular dynamics simulations to sample the transition path distribution, we will generate in an iterative

way short trajectories. This iterative procedure will be a random markovian process that only depend on the previously generated trajectory. If the procedure is carried for long enough, it will reach a stationary state that is exactly the transition path distribution [15]. Clearly in practical situations it can be far from trivial to assess whether such stationary converged behavior is reached or not. The stopping criterion is choosed case by case and is generally not independent to the computer time limitations. It is important to remark that a similar concern and difficulty in providing clear convergence indicators and – related to this – reliable statistical error bars, is common to all sampling techniques based on molecular dynamics or Monte-Carlo methods.

Among the different sampling strategies, transition path shooting emerged as particularly efficient. The general algorithm work as follow : a shooting point along an existing reactive trajectory is chosen at random, its velocities are perturbed, and then from this point trajectories are propagated forward and backward in time [41]. Practically this means that one is launched with inverted velocities to achieve inversion of time. If the two segments end up in different states, the new path is accepted, otherwise it is rejected. To sample all the transition path we repeat this procedure iteratively, as illustrated in figure 5.2.1.

About the efficiency

In theory all the transition path shooting algorithm are based on rigorous mathematical equation and should sample an exact path ensemble. In practice however, there are two main limitations to achieve this sampling in a reasonable amount of time and machine power. The first one is linked to the typical duration of the relaxation from the shooting point. As an extreme example, if the system requires several microseconds to relax toward one of the two states, it would take years and dozens of millions of cpu hours to have a proper sample. The second one is linked to the decorrelation of the sampled transition path. Again as an extreme example, if it takes millions of steps to have two decorrelated path, that is to sample two truly distinct pathways, if would render the sampling impossible to carry.

So to have efficient transition path shooting, we need that typical generated path are short and that they decorrelate after a few iterations of the algorithm. This requires that the trials path are accepted with a reasonably high ratio, that we will call the acceptance rate in the following. This also requires that every part of the path ensemble is reachable by the iterative sampling procedure. Sadly the first conditions is rarely met in real case uses, especially in material science. The probability to create a transition path TP starting from a phase point $q = (x, v)$ is

$$p(TP|q) = \phi_A(q)\phi_B(\underline{q}) + \phi_B(q)\phi_A(\underline{q}) \tag{5.2.1}$$

where ϕ_B and ϕ_A are the committor functions and $\underline{q} = (x, -v)$. As discussed in (4.2.5), for typical free energy barrier met in condensed matter, the committor will be close to 0 or 1 everywhere, except in a small transition region. Therefore product of the two $\phi_A\phi_B$ will be near zero everywhere, except on the transition region, that is the top of the free energy barrier [100, 101, 41].

Thus it should be no surprise that we can address this issue by building markovian processes that pick shooting point only in the transition state domain. This is exactly

what the aimless shooting and its derivative the aimless shooting within a range algorithm are doing. Before detailing them and their differences, we will briefly present the seeding method used to generate the initial reactive trajectories.

5.2.1 Seeding

Seeding is a method that allows one to easily generate initial reactive trajectories, and to compute various thermodynamic and kinetic properties when applied to nucleation. It is not a true transition path sampling method, as it is mostly an empirical method based on clever tricks, without any rigorous mathematical background to back-it up.

It was specifically developed to tackle the discrepancy in time scale when studying ice nucleation. We will present it more in detail later in chapter 7, but nucleation processes have two different time scale : the first one is linked to the probability to see appearance of a sufficiently large initial nucleus and is generally very long (seconds or microseconds at best). The second one is linked to the growth speed of ice, which can be very fast (few hundreds of nanoseconds). Hence if one wanted to study for instance the critical nuclei size, the size from which nucleus will "certainly" expand themselves to the whole system, it would requires molecular dynamics simulations of thousands of microseconds or seconds to get relevant statistics, which is unimaginable nowadays, even though the interesting events themselves would have really short duration.

Seeding solves this issue with a simple procedure

— generate from its unit cell a perfectly crystalline box of the ice phase under study ;
— extract from this box an initial nucleus with a specific shape, usually a sphere or a slab ;
— solvate this nucleus with liquid species and equilibrate the interface ;
— from this initial seed, generate several molecular dynamics simulation, each with random new momenta drawn from the Boltzmann distribution at the relevant temperature.

As said, contrary to the apparition of a sufficiently large nucleus, the nucleus grows in a small time scale, of the order of the nanoseconds, that is perfectly reachable for today simulations. By repeating this procedure, one can evaluate for various sizes the committor functions and so get an estimation of the critical nuclei size, where $\phi_A \approx \phi_B$. Using classical nucleation theory, that will be presented in section (7.1.1), one can estimate the free energy barrier and the nucleation rate from knowledge of the critical nucleus size, the attachment rate and the free-energy difference between liquid and solid bulk phases [102, 33, 35].

So it seems that seeding is the perfect technique to study nucleation : it's cheap, easy to implement and may be used in combination with other techniques to give access to the most relevant information about thermodynamic (free energy) and kinetic (nucleation rate) properties of the system. However in practice this methods have two flaws. First one is that it relies heavily on classical nucleation theory, which is too simplistic for some materials. Second one is that this methods is heavily dependent on the initial choice of the nucleus structure, and so on the nucleation pathway, which can be far off from the real processes that happen in nature. Hence it generally give rough estimates of the previously mentioned quantities, working best with large critical nuclei size [34].

These limitations in head, in this thesis we will mainly use seeding to generate initial reactive trajectories, before using more advanced and rigorous enhanced sampling technique presented here. We will present the precise procedure to produce the initial seed in (7.2.2).

We want to stress that despite its flaws, seeding has been applied to extract a large amount of important informations with realistic potential for water, including critical nucleus size, speed of growth or melting of the nucleus, free energy barrier and nucleation rate, as discussed in section (7.1.4). Furthermore, these investigations opened the path for a large amount of studies based on more rigorous methods, since they allow to choose sufficient simulation box sizes, simulation duration, and optimal temperature. In our study, indeed, we followed this path.

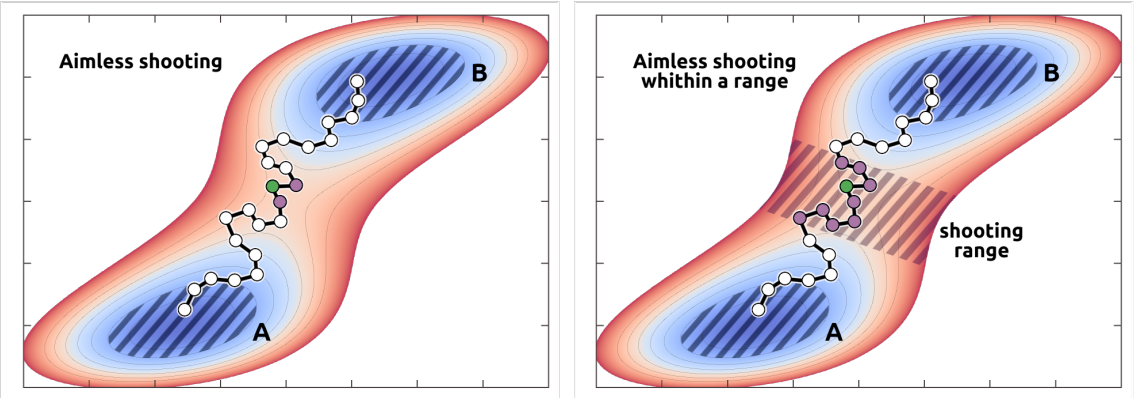


FIGURE 5.2.2 – Difference between the aimless shooting and aimless shooting within a range. Green point represent the last accepted shooting point, purple points represent the potential new trial point. For aimless shooting on the left, only two new point can be picked in the direct vicinity of the last accepted point. For aimless shooting within a range, all points that fall within a predefined range, here represented by dashed lines, can be chosen, including the last accepted point.

5.2.2 Aimless shooting algorithm

The first transition path shooting procedure that we used is aimless shooting [12, 13]. The strength of this algorithm is that it is able to concentrate shooting point attempts without the use of an order parameter to describe the transition region. In fact it still requires an order parameter to distinguish the two states and to define their basins, which in practice is relatively easy to find and a less stringent requirement.

Aimless shooting achieve this feats in a simple and clever way, by only drawing trial shooting point in the nearest vicinity of the last accepted one. This however is also its major drawback, as it will produce sequence of transition path that tend to pass through a narrow region and share a common point. Which means that it will typically take a large amount of step to achieve decorrelation of the sampled transition path.

Starting from an initial reactive trajectory with a maximum duration T, aimless shooting consists of iteration of these steps :

1. discretize last accepted trajectory in a series of candidate shooting points with a time step $\Delta t \ll T$;

2. if the last shooting point was at frame i, select randomly with equal probability the frame $i + \Delta t$ or $i - \Delta t$ as the new shooting point ;
3. draw new momenta from the Boltzmann distribution and propagate the system backward and forward in time, until the trajectories reach one of the two states or reach maximum duration ;
4. accept the new trajectory if it connects the states A and B.

Due to its way of randomly choosing the new trial shooting point, the acceptance rate of the algorithm is correlated to the choice of the time step Δt chosen to discretize the trajectory. But this time step is also linked to the decorrelation of two consecutive transition path. When $\Delta t \to 0$, the acceptance rate r will tend toward 100%, but in the same time two consecutive transition path sampled will be highly correlated. On the reverse, if Δt becomes similar to T, $r \to 0$ but we will have high decorrelation between two consecutive transition path sampled. Generally one will chooses $10\% < r < 25\%$ to have a good trade off between the two factors.

As a general remark for this type of methods, it can be non trivial to identify a criterion for optimal choice of parameters. Notwithstanding the importance to have a non-null acceptance rate, practical ways to assess the efficient evolution from an unlikely transition state region to the optimal one, and the decorrelation speed between different transition paths should be identified. For instance in nucleation, to have a relatively high acceptance rate is not sufficient to have quick evolution of the critical nucleus structure, see section (7.4.3).

Due to its theoretical simplicity and relative ease of implementation, aimless shooting has been employed to study several systems and it has been especially useful in the study of nucleation [103, 104, 38, 105].

5.2.3 Shooting range algorithm

The second transition path shooting procedure that we used is shooting range [41]. The algorithm was specifically devised to fix the decorrelation issue of the aimless shooting algorithm, while still achieving high acceptance rate.

Its main drawback compared to aimless shooting is that this algorithm requires a reaction coordinates able to describes correctly the transition region. In fact its main difference with aimless shooting is that instead of drawing trial shooting points near the last accepted one, we will draw them in a whole shooting range that span the transition region.

Explicitly the algorithm is as follow :

1. discretize last accepted trajectory in a series of candidate shooting point with a time step $\Delta t \ll T$
2. select randomly with equiprobability the new shooting point from the ones that are within the predefined shooting range ;
3. draw new momenta from the Boltzmann distribution and propagate the system backward and forward in time, until the trajectories reach one of the two states or reach maximum duration ;
4. accept the new trajectory if the trajectory connect liquid and ice *and* with the following acceptance probability $p_{acc} = \min(1, n/n')$. Where n and n' are the number of points of the trajectory that are within the shooting range, for old and new trajectory respectively.

Obviously the efficiency of the algorithm is dependent of the shooting range. It is possible to use a systematic procedure to optimize this choice [41], but it requires a significant amount of "spoiled" steps. So here we will resort to more empirical methods to choose the shooting range, based on knowledge acquired with seeding, as will be described in section (7.4.2). The difference between the two way of picking randomly new shooting points is summarized in figure 5.2.2.

We want to stress that this algorithm is relatively new, and so far it has only been applied to a few systems [106]. We applied it for the first time to the challenging problem of homogeneous ice nucleation of water, finding that if one has access to a good reaction coordinates and with a correct choice of shooting range, for system where the transition region span a wide range of possible states it is extremely efficient. Notably this algorithm is incredibly faster than aimless shooting to achieve decorrelation of transition path sampled, allowing fast study of the evolution of critical nuclei structures, where two different symmetry are in competition, see section (7.4.3) and (7.4.4).

Deuxième partie

Results

6 Study of the Liquid-Liquid Phase Transition

6.1 Introduction

Among the many peculiarities and anomalies of water, several experiments have disclosed connections between stable crystalline phases and metastable amorphous phases [107]. The first found was between the stable crystalline ice at ambient pressure (Ice I), and the low-density amorphous (LDA) and high-density amorphous (HDA) ices : by compressing Ice I up to 10 kbar at $\approx 80K$, one obtains the high-density amorphous ice instead of the stable Ice VI [17]. This high-density amorhous can be transformed into low-density amorphous by decompression at 130 K [16]. Finally one can recover Ice I by heating up this low-density amorphous [108]. Similar connection can be found between crystalline and amorphous ices in the high-pressure region of the water phase diagram, where connection between a very-high-density amorphous (VHDA) ice [18], plastic ice and the stable Ice VII have been observed or predicted [109, 110].

This properties of water to possess several amorphous phases at low temperature, namely LDA, HDA and VHDA, is called polyamorphim. It has clearly been one of the most puzzling anomalies of water in the last decade, as apparently reversible first-order transitions among some of them [19], seems in fact at odds with the very thermodynamic notion of metastable glassy forms.

Several scenarios have been formulated to explain these phenomena, the most famous, and somehow controversial, being the occurrence of a first-order liquid-liquid transition in supercooled water, extending at lower temperatures in the amorphous region, and terminating with a second critical point at higher temperatures. This hypothesis was formulated on the basis of a computational molecular dynamics study, using the ST2 model of water [20]. Precisely, it used small box of 216 ST2 water molecules to compute the density at several conditions of pressure and temperature, using similarities with the LDA-HDA transformation to hypothesized a liquid-liquid transition.

It is however extremely challenging to verify this hypothesis with experiments in pure bulk water. Indeed, its supposed location would lie below the kinetic limit of homogeneous ice formation, in the so called *no man's land* [44, 110, 111, 1]. That is, a region in which spontaneous nucleation in the stable Ice I phase occurs, regardless of if we cool down liquid or heat up amorphous ice, preventing any observation of the underlying metastable liquid phase that may exist in it.

6.1.1 The two-states model

Important theoretical efforts have since been made, in order to improve our understanding of polyamorphism, notably in the liquid phase. For example, the definition of a two-states model provides a unitary description of the thermodynamics for most polymorphic fluids [45]. In this view, water is considered as a "mixture" of two interconvertible local structures : a high-density, high-entropy liquid and a low-density, low-entropy liquid [112, 113, 114, 115]. The model predict four scenarios, discriminated through the density extrema loci [45] :

— a singularity-free scenario, with interconversion between two states but no phase separation ;

— a liquid-liquid critical point scenario, with interconversion and phase separation ;
— a degenerate case where the critical point coincides with the vapor-liquid spinodal ;
— a critical-point-free scenario, with a virtual critical point located below the vapor-liquid spinodal.

This model is elegant, and its predictions intriguing, but critically dependent on the detailed choice of the thermodynamic parameters.

6.1.2 Experimental studies of the liquid-liquid transition

From the experimental point of view, a huge battery of diverse set-ups have been deployed, over the years, using for example aqueous solutions, or confined/micro-sized systems, in order to overcome the thermodynamic frontiers of the no man's land, while avoiding the inevitable crystallization of supercooled water into ice [44]. Some of those experiments have been able to firmly establish that the transition between the low and high density amorphous ices is a first order one [19, 116, 117, 53, 118, 119]. Some experiments were able to give hints on the presence of two competing liquid forms in the supercooled region [51, 52, 53, 54, 55]. For bulk water, recent experiments carried out at negative pressures suggest that the right scenarios are either the singularity free or the liquid-liquid critical point [120]. Other experiments conducted on salty water, give instead strong arguments against a first order liquid-liquid transition in the supercooled region [21], even if a first order transition were observed for the corresponding amorphous phases [22], hence showing that no direct link necessarily exists between polyamorphism and a liquid-liquid transition.

6.1.3 Numerical studies of the liquid-liquid transition

From the computational point of view, the second critical point scenario has been a long source of debate since its very first proposition [20], mostly because that work was based on the "ST2 model" of water [4], which is known to be significantly overstructured, and thus to "enhance" certain anomalies of water. After several free-energy studies found contradicting results with this model, either demonstrating the LDL-HDL transition and coexistence [81, 121, 122], in systems containing up to 600 ST2 water molecules [23], or supporting a no-transition scenario with up to 512 molecules [123, 124]. A consensus emerged on the former hypothesis, thus validating phase coexistence and reconciling the two independent free energy calculations [125, 1]. However, this result seems limited to this specific model, nowadays known for its drawbacks, and widely considered as not particularly representative of real water, see section (3.2.2).

Other studies pointed out in fact that the thermodynamics of the putative LDL-HDL transition in supercooled water was heavily model-dependent [126, 127, 76]. In the last few years, the so-called TIP4P/2005 force field [5] has emerged as one of the most accurate models, as it reproduces quite accurately the phase diagram and anomalies of water [73, 72, 74, 75], as discussed in section (3.2.3). Several numerical studies were performed with TIP4P/2005 to assess the existence of a liquid-liquid transition, although none with a thorough and extensive free-energy approach. A critical point for TIP4P/2005 water was first proposed at 1.35 kbar, 193 K and

1012 kg. m^{-3}, based on the analysis of density and concentration fluctuations in the supercooled region in 500-molecules models for durations of 500 ns [128]. A subsequent analysis failed to reproduce this result with larger boxes and longer simulations (1,000 to 32,000 molecules and 500 ns to 5 μs) and showed that size effects are important, together with the long relaxation time of the system [56], as it was confirmed later on [129, 130]. Another study looked at density fluctuations concluding that they constitute the signature of a liquid-liquid transition [131], but once again a subsequent analysis with larger simulation boxes argued that their origin is the appearance of ice-like structure [57]. With the coupled use of longer simulations and a two-state thermodynamic analysis, a new critical point was proposed at 182 K and 1.70 kbar [24, 25], consistently with previous numerical [56] and experimental studies [120].

More recently another study was published, based on the analysis of density fluctuations [26]. The authors combined extensive unbiased simulations of tens of μs for 300, 500 and 1000 molecules, at $T \geq 177$ K for TIP4P/2005 and ≥ 188 K for TIP4P/ice, i.e., above the postulated second critical point of the two models (see below), with an histogram reweighting technique to extrapolate order parameter distributions at lower temperature, closer to the supposed critical regime. By fitting the extrapolated distributions, together with static scattering functions computed on larger boxes at $T > 180$ K, to a 3D Ising model, an estimation of the liquid-liquid critical point conditions is obtained at $T_c = 172 \pm 1$ K and $P_c = 1861 \pm 9$ bar for TIP4P/2005. This elegant work still is not a proof of the liquid-liquid phase transition (LLPT), as the authors themselves write "*Rigorous proof of the existence of a LLPT requires performing free energy calculations at subcritical temperatures.*".

The present work aims precisely at this much-needed rigorous proof, by overcoming several issues emerged from the large corpus of computational studies carried out over the last 30 years, and at providing a robust answer to this long-going question. To this end, we adopt a strategy based on several methodological strengths. First, we employ a versatile topological metric to describe structural transformations in water, already proved to be very effective in discriminating the known crystalline, amorphous, and liquid forms of water [84], and that we successfully used to study several phase transitions throughout the phase diagram of water, including the extremely challenging spontaneous nucleation of crystalline ice from the bulk liquid [8]. Second, we exploit a synergistic free-energy calculation approach, combining metadynamics to explore the configuration space, umbrella-sampling to collect extensive statistics along the transformation paths, and unbiased MD trajectories probing the spontaneous evolution from different phase-space regions to validate free energies and extract valuable dynamic information. Third, we use the TIP4P/2005 force field, which is nowadays considered the most reliable and accurate to describe real water. We fully describe our approach in the Materials and Methods section.

Anticipating our results, the combination of these advanced techniques and demanding calculations allows us to establish the relative "flatness" of the free-energy landscapes throughout the no man's land, and thus to suggest that no LDL-HDL first-order transition exists in the supercooled regime, differently from what is experimentally observed in the amorphous region.

6.2 Simulation methods

6.2.1 Molecular dynamics parameters

We performed molecular dynamics simulations employing the TIP4P/2005 [5] inter-atomic potential, for all the reasons presented in (3.2.3), with periodically-repeated triclinic boxes containing $N = 800$ water molecules. Over the course of the simulation the size and shape of the box will vary slightly, the average box vector are represented in table 6.2.1.

P, T	$\langle A \rangle$ (Å)	$\langle B \rangle$ (Å)	$\langle C \rangle$ (Å)	$\langle \alpha \rangle$	$\langle \beta \rangle$	$\langle \gamma \rangle$
160 K, 2.5 kbar	24.7 ± 0.1	23.2 ± 0.1	42.3 ± 0.2	100	79	110
170 K, 2 kbar	24.6 ± 0.1	23.2 ± 0.1	43.0 ± 0.2	100	79	109
180 K, 2 kbar	22.8 ± 0.1	28.1 ± 0.1	38.0 ± 0.1	96	83	101
P, T	$\langle A \rangle$ (Å)	$\langle B \rangle$ (Å)	$\langle C \rangle$ (Å)	$\langle \alpha \rangle$	$\langle \beta \rangle$	$\langle \gamma \rangle$
160 K, 2.5 kbar	24.0 ± 0.1	23.6 ± 0.1	38.9 ± 0.1	96	87	106
170 K, 2 kbar	24.7 ± 0.1	23.8 ± 0.1	39.5 ± 0.1	95	83	107
180 K, 2 kbar	22.3 ± 0.1	28.6 ± 0.1	36.4 ± 0.1	89	95	109

TABLE 6.2.1 – Average simulation box parameters for selected P, T conditions computed from unbiased shooting trajectories. The first half of the trajectory is discarded as equilibration. The two columns present average values from trajectories starting from low ($S = 1.02$, top) or high-density ($S = 1.98$, bottom) states.

All simulation were done under NPT conditions between $140 - 182$ K and $1 - 5$ kbar, employing the gromacs 5.1.4 simulation package [132]. We adopted a 2 fs timestep. Short-range interactions were truncated at 0.85 nm, and the particle mesh Ewald method was used to compute electrostatic interactions. Bond constraints were maintained using the LINCS algorithm with a fourth order expansion [133].

To control the temperature we used the stochastic velocity rescaling thermostat with a relaxation time of 0.5 ps [65]. At first for the pressure we used an isotropic Berendsen barostat with a relaxation time of 0.5 ps [66]. But as discussed in (3.1.4), this barostat does not yield the correct NPT ensemble, leading to potential un-physical density fluctuations. Thus we switched to the correct Parrinello-Rahman barostat [67], as will be discussed in section (6.3.1) and (6.3.2).

6.2.2 States preparation

To study the properties of the no man's land region, we first needed to prepare a series of sample states at various condition of pressure and temperature. To do that we first generated with gromacs a liquid box of 800 molecules. Then we equilibrated it first in NVT ensemble at 180 K for 5 ns, and then in NPT ensemble at 0 bar for 5 ns.

From this initial state at 180 K and 0 bar, we generated a series of states ranging from 170 to 140 K by cooling down our system by steps of 10 K, doing short equilibration of 5 ns each time. That is we first goes down to 170 K, equilibrate for 5 ns, then switch to 160 K, equilibrate for 5 ns... and repeat until reaching 140 K.

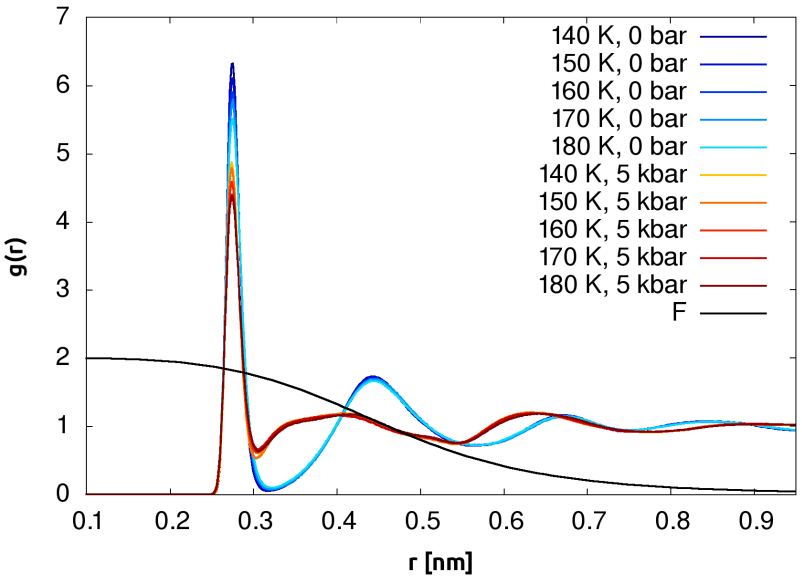


FIGURE 6.2.1 – Radial distribution functions g of the low and high-density reference states ($P = 0$ or 5 kbar respectively) at different temperatures, with the switching function $\mathsf{F} = 2\sigma$ used to define PIV, multiplied by 2 to ease the visualization.

Then for each temperature, we performed a compression cycle. We increased the pressure by step of 0.5 kbar, each time equilibrating for 5 ns, until reaching 5 kbar. Here we are not performing a perfect structural equilibration at each pressure, but is is enough to initialize the metadynamic simulations that we will use to explore the configuration space available. Also this compression cycle furnish us with correct low and high-density liquid at 0 and 5 kbar respectively.

6.2.3 Order parameter definition

As already said one of the aim of this thesis was to assess the quality and usefulness of PIV with various system. Thus it should be no surprise that we will use it to define an order parameter able to distinguish low and high-density liquid configurations.

Here we coupled the PIV distance with the path collective variable, see (4.2.3) and (4.2.4) for their respective definition. Here to define the PIV we only used direct pair of atoms, with Oxygen-Oxygen and Hydrogen-Hydrogen distances, which lead to two PIV block

$$V_{ij}^1 = w_1 \sigma \left(\left(\frac{V}{V_0} \right)^{1/3} r_{ij}^{OO} \right) \tag{6.2.1}$$

$$V_{ij}^2 = w_2 \sigma \left(\left(\frac{V}{V_0} \right)^{1/3} r_{ij}^{HH} \right) \tag{6.2.2}$$

where $r_{ij}^{\alpha\alpha}$ is the distance between atom i and j of type $\alpha \in O, H$. The reference volume is $V_0 = 0.024$ nm^3. The PIV blocks were weighted with $w_1 = 1$ and $w_2 = 0.2$. The switching function used is a rational one with the following formula

$$\sigma(r) = \frac{1 - (r/r_0)^4}{1 - (r/r_0)^{10}} \tag{6.2.3}$$

with $r_0 = 0.5$ nm. This specific choice of switching function was made to have maximum variation between the first and second coordination shell, including their $g(r)$ peaks as shown in figure 6.2.1.

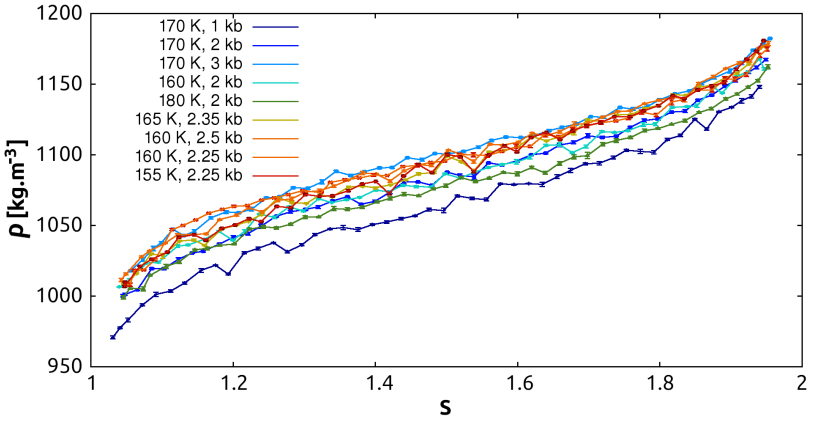


FIGURE 6.2.2 – Average density as a function of the S path coordinate for every umbrella sampling window for various P, T conditions. An almost linear correlation can be observed. The bars indicate the standard deviation of the density.

To define the path collective variables, we used the previously equilibrated structure at 0 and 5 kbar for each temperature as our two references. That is if we perform a simulation at 170 K, we will use the liquid equilibrated at 170 K and 0 or 5 kbar. Noting H and L these high and low-density structure and using the PIV distance as our metric, we can define S and Z simply

$$S(X) = \frac{1 \times \mathrm{e}^{-\lambda D_{LX}} + 2 \times \mathrm{e}^{-\lambda D_{HX}}}{\mathrm{e}^{-\lambda D_{LX}} + \mathrm{e}^{-\lambda D_{HX}}} \tag{6.2.4}$$

$$Z(X) = -\frac{1}{\lambda} \log\left(\mathrm{e}^{-\lambda D_{LX}} + \mathrm{e}^{-\lambda D_{HX}}\right) \tag{6.2.5}$$

where D_{LX} is the PIV distance between low-density structure and a configuration X, and D_{HX} is the same for high-density structure. Note that we are speaking about structure and not liquid, as at 140 K we are entering in the amorphous domain. We chose $\lambda = 0.3$, following the common rule of thumbs that λ times D_{LH} should be equal to 2.3.

We want to stress again that this way of defining an order parameter S using PIV is very general, as it can be applied to transitions between ordered or disordered structures in different materials [8]. Also as you may see it requires In the specific case of the liquid-liquid transition S is highly correlated with the density of the system, as shown in figure 6.2.2, which is known to be a good order parameter for this transition [123, 124, 23].

6.2.4 Free energy calculations

With our order parameter defined, we were ready to start computation of free energy profile at selected P, T conditions. For each point we performed enhanced-sampling simulations aimed at reconstructing the free-energy landscape for the supercooled liquid, using the open-source, community-developed PLUMED library version 2.6 [79]. The procedure was in two steps : first, we exploited metadynamics to

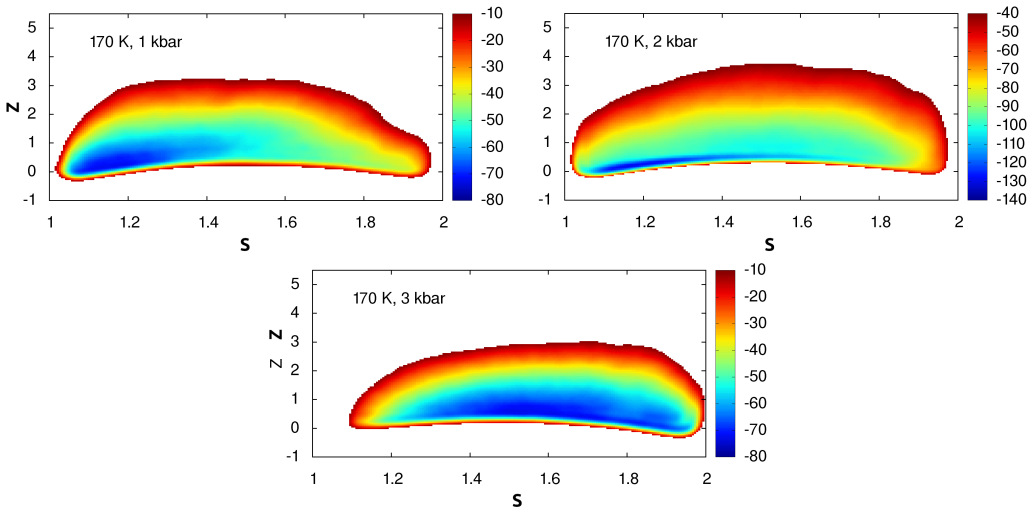


FIGURE 6.2.3 – Bias surface reconstructed from metadynamics at 170 K and 1, 2 and 3 kbar as a function of the PIV-based path coordinates S and Z. The color scale is in kT units. No significant feature can be observed along the Z direction.

obtain transition pathways as well as a preliminary estimate of the free energy landscapes [39]. Then, we reconstructed statistically converged free energy profiles with more expensive umbrella sampling simulations [14].

For metadynamic, we have done simulations of 25 ns to 50 ns, placing gaussian hills of width $\sigma_S = 0.015$, $\sigma_Z = 0.15$ and height of 0.239 kcal/mol every ns. During this simulations the system easily pass from one liquid state to another several times. As shown in figure 6.2.3 for three set of pressure at 170 K, the estimated free energy profile have no specific features along the Z coordinates. A property that is the same in every (P, T) conditions analyzed in this work.

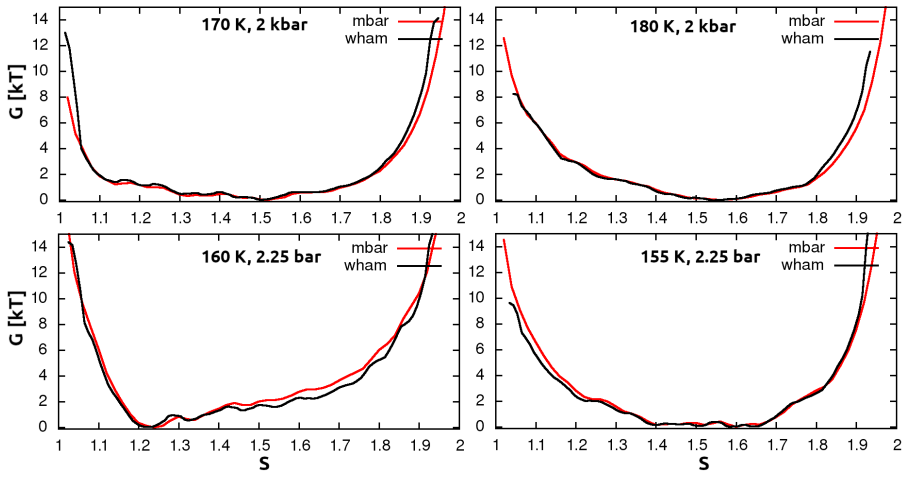


FIGURE 6.2.4 – Comparison of two techniques used to compute free energy profiles from umbrella sampling simulations : in red the Multistate Bennett Acceptance Ratio (MBAR) or binless Weighted Histogram Analysis Method (binless-WHAM) from Ref. [92, 134], and in black the WHAM from Ref. [91, 135]. The first quarter of each trajectory was discarded as equilibration.

Using the result of metadynamics, we could resort to more efficient and controllable umbrella sampling simulations. We only sampled the S coordinates as no important information of $G(S, Z)$ were along Z. Starting from configurations explored with metadynamics, we used 48 windows spaced by $\Delta_S = 0.02$, ranging from $S = 1.02$ to $S = 1.98$. The harmonic bias potential defined in equation 5.1.1 had a spring constant $\kappa = 2826.5$ kcal/mol. The length of the simulations are dependent of the P, T conditions and will be precised in section (6.3.3).

Finally, the data accumulated in the different windows were combined together to compute the free energy profile by means of the binless weighted histogram analysis method (called also multistate Bennett acceptance ratio) [92], using open source code from Joshua Goings (https ://github.com/jjgoings/wham). For comparison, we also reconstructed free-energy profiles using Alan Grossfield's implentation [135] of the traditional method in Ref. [91]. Anticipating a bit, figure 6.2.4 shows that there are at most 1 $k_{\mathrm{B}}T$ of difference between the two methods. Convergence and error estimation will be discussed deeply in section (6.3.3).

6.3 Exploration of the (P, T) diagram

With the initial states previously equilibrated and the free energy calculation methods described previously, we can now freely start to sample the free energy at various pressure and temperature conditions, to see what the phase diagram look like and if there is or not a first order transition.

6.3.1 First exploration using Berendsen barostat

To do that we followed an iterative process, where we computed free energy landscape G of one point in (P, T) space, and then from the knowledge acquired we moved onto another to explore the no man's land in a significant but not too costly way, using short duration of 15 ns for umbrella sampling.

The first point computed was at 180 K and 2 kbar, where we found that G has one minima not localized around a clear low or high density liquid, with a broad range of value within few $k_{\mathrm{B}}T$. As we wanted to assess the effect of temperature on the stability of the two state, so we kept the same pressure and computed point from 170 K down to 140 K by step of 10 Kelvin. Figure 6.3.1.a shows that for 170 K at 2 kbar, G was mostly flat with a minima localized around the low density liquid state. For all the other G was clearly leaned in favor of the low density liquid.

As a general guide to interpret the S-space, low-density water features $S \lesssim 1.5$, and the opposite for high-density water. Next, we wanted to see see how G evolved with the pressure, we fixed the temperature at 170 K and then computed point from 1 up to 5 kbar by step of 1 kbar (obviously skipping the already compute 2 kbar point). Figure 6.3.1.b shows that for low pressure low density liquid is more stable, and that for pressure higher than 2 kbar its the high density liquid.

After this first exploration, we took several P, T with intermediate conditions compared to those that favor low or high density states. Figure 6.3.1.c shows that it give us a line of relatively flat free energy profile, where both low and high density state are mostly equiprobable.

Even if all these calculation give us valuable knowledge, they were all performed with the Berendsen barostat, which does not yield a correct NPT ensemble and may

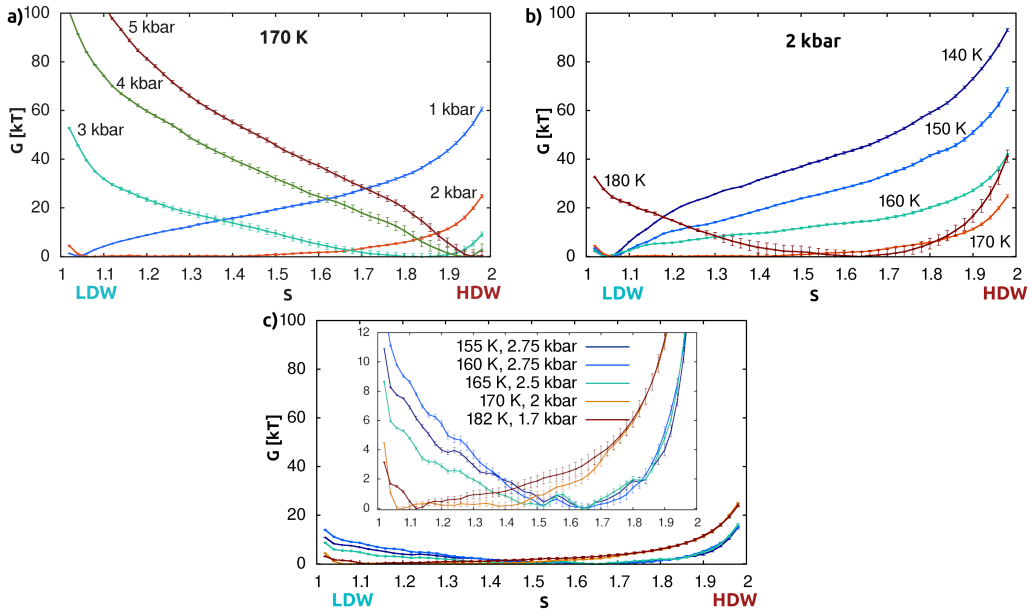


FIGURE 6.3.1 – Free-energy profiles for the low-density/high-density water transformation with incorrect use of the ensemble inconsistent Berendsen barostat. $S \approx 1.1$ correspond to low density and $S \approx 1.9$ to high density. (a) Temperature ranging from 140 to 180 K along a 2 kbar isobar. (b) Pressure ranging from 1 to 5 kbar along a 170 K isotherm. (c) Conditions intermediate between those favoring low density and high density ; note the relatively flat free-energy profiles (see the zoomed inset).

lead to unphysical fluctuation of volume or density in a non-trivial way, see section (3.1.4). To test this, we computed G for 182 K and 1.7 kbar for both barostat. As we found less than 1 kT of difference between the two, we thought that despite its flaws the Berendsen barostat worked in a correct way for this specific problem. But a reviewer pointed out to us that even if ensemble inconsistent sampling may give correct results for some specific conditions of pressure and temperature, they still may fail in unpredictable ways for others, mentioning the controversy around the presence of a free energy barrier in the ST2 model.

6.3.2 Proper exploration with ensemble consistent barostat

This is why we switched to the Parinnello-Rahman barostat, which yield a correct NPT ensemble [62]. First we recomputed points along the 2 kbar isobar or the 170 K isotherm, to see how the phase stability evolved under the change of barostat. This time we avoided the high density (> 3 kbar) and the low temperature (< 160 K), as we knew from previous calculation that these points displayed no specific features.

Figure 6.3.2.a shows that the change of barostat does not incur a profound change of the physical property of the supercooled liquid, as we have still monotonous free energy profile. For both barostat, by following the 170 K isotherm, or the 2 kbar isobar, the free-energy profiles always exhibit a single minimum along the transformation path. This free energy minimum move from low to high S values along the isotherms when we increase the pressure, which correspond to a continuous switch of stability between the low and high-density liquid.

In fact, it seems that the main effect of the correct ensemble sampling is to

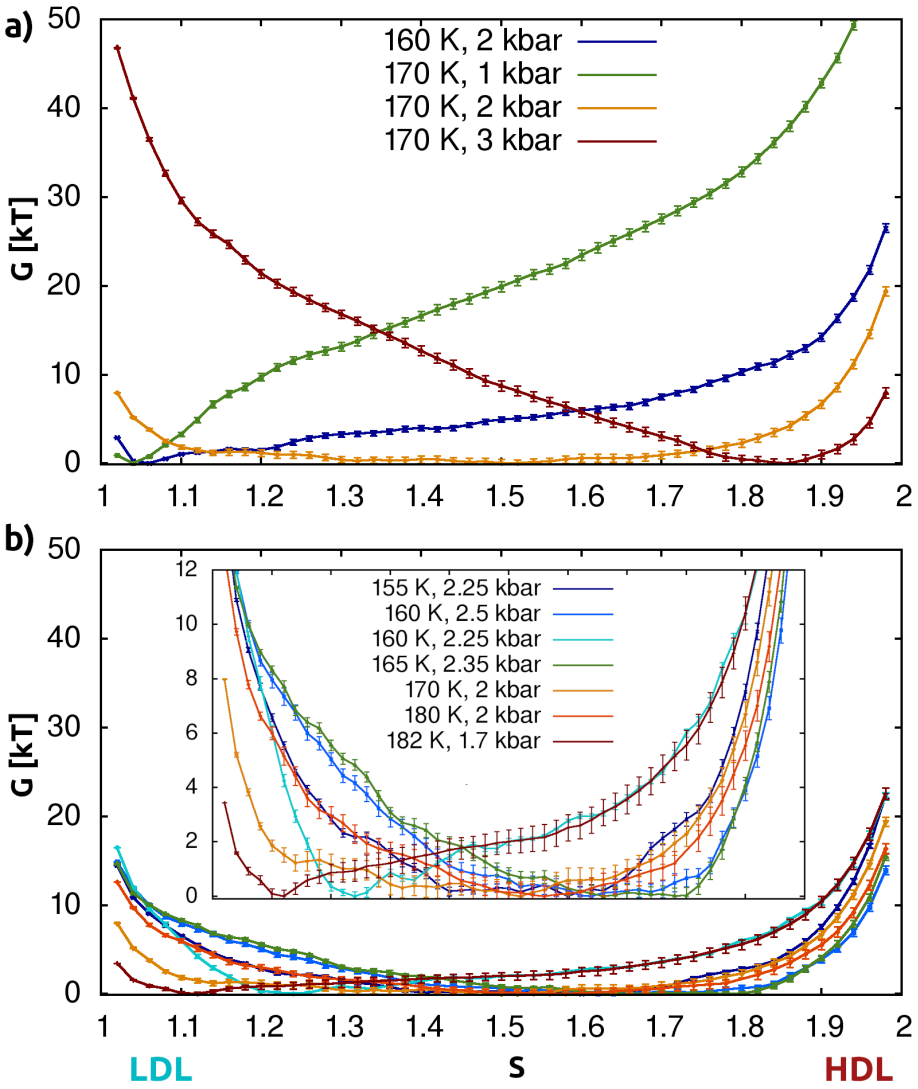


FIGURE 6.3.2 – Free-energy profiles for the low-density/high-density liquid transformation with . $S \approx 1.1$ correspond to low density and $S \approx 1.9$ to high density. (a) Pressures and temperature around 170 K and 2 kbar. (b) Conditions intermediate between those favoring low density and high density ; note the relatively flat free-energy profiles (see the zoomed inset).

augment the stability region of the high density liquid, by reducing the required pressure to make it stable, as can be seen by the flattening of the free energy profile at 160, 170 K and 180 K at 2 kbar. As for the Berendsen barostat, a natural question arises : what is the precise shape of the free-energy landscape at conditions where low and high-density water forms are equiprobable ?

Remarkably, for these (P, T) conditions that are intermediate with respect to those favoring low or high density, we still observe relatively flat free energy profiles (within a few $k_{\mathrm{B}}T$ units), without any sizable barrier separating low and high-density liquid, as shown in figure 6.3.2.b. Such flat profiles indicate that the system populates a relatively broad range of different densities and coordination numbers, as can indeed be observed in figure 6.3.6.b, and 6.3.7a.

6.3.3 Convergence assessment and error estimation

Contrary to the first exploration with Berendsen barostat which used short duration of 15 ns for the umbrella sampling, this time we used much longer sampling duration. Precisely, the length of the simulation is dependent of the P, T conditions : for 160, 180 K at 2 kbar and 182 K at 1.7 kbar the simulations were 25 ns long. For 170 K at 1 and 3 kbar, the simulations were 50 ns long. For 170 K at 2 kbar the simulation was 60 ns long. For all other simulations, they were 100 ns long.

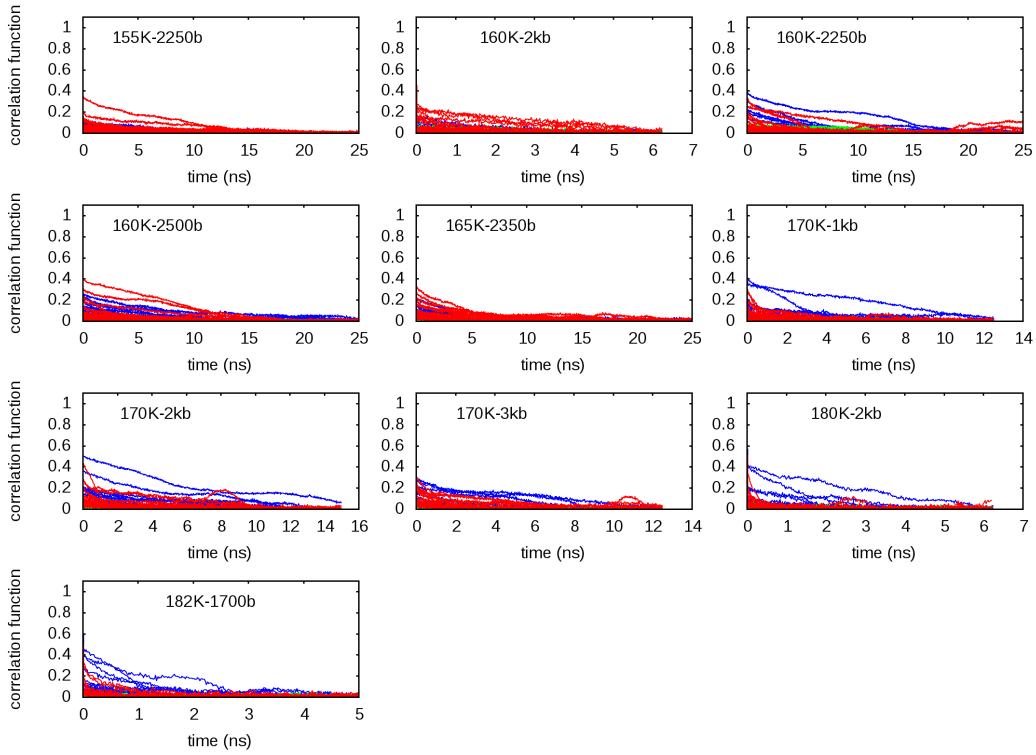


FIGURE 6.3.3 – Normalized self-correlation function $C_i(S(t))$ of the PIV-based path coordinate S in umbrella sampling simulations. Each line corresponds to an umbrella sampling window, with $S < 1.4$ in blue, $1.4 < S < 1.6$ in green, and $S > 1.6$ in red. The first fourth of each trajectory is not employed to compute the correlation, and the total length of each trajectory is four times the duration plotted.

As a first indication of the convergence of umbrella sampling, we have computed the auto-correlation function of the S-path coordinate in each umbrella sampling windows :

$$C_i(s(t)) = \frac{\langle \delta s_i(0) \delta s_i(t) \rangle}{\langle \delta s_i^2 \rangle} \qquad \text{with} \qquad \delta s(t) = s(t) - \langle s \rangle \tag{6.3.1}$$

discarding the first quarter of each trajectory as equilibration. Figure 6.3.3 reports these auto-correlation functions, and we also report for comparison the corresponding functions for the density ρ and the sixth-order Steinhardt parameter $\langle Q_6 \rangle$ in figure 6.3.4 and 6.3.5, respectively. The important thing here is that they are all significantly shorter than the duration of their related umbrella sampling windows.

As a second indication, we estimated statistical uncertainties on free-energy profiles using block averages, taking the largest value of the standard error of the mean.

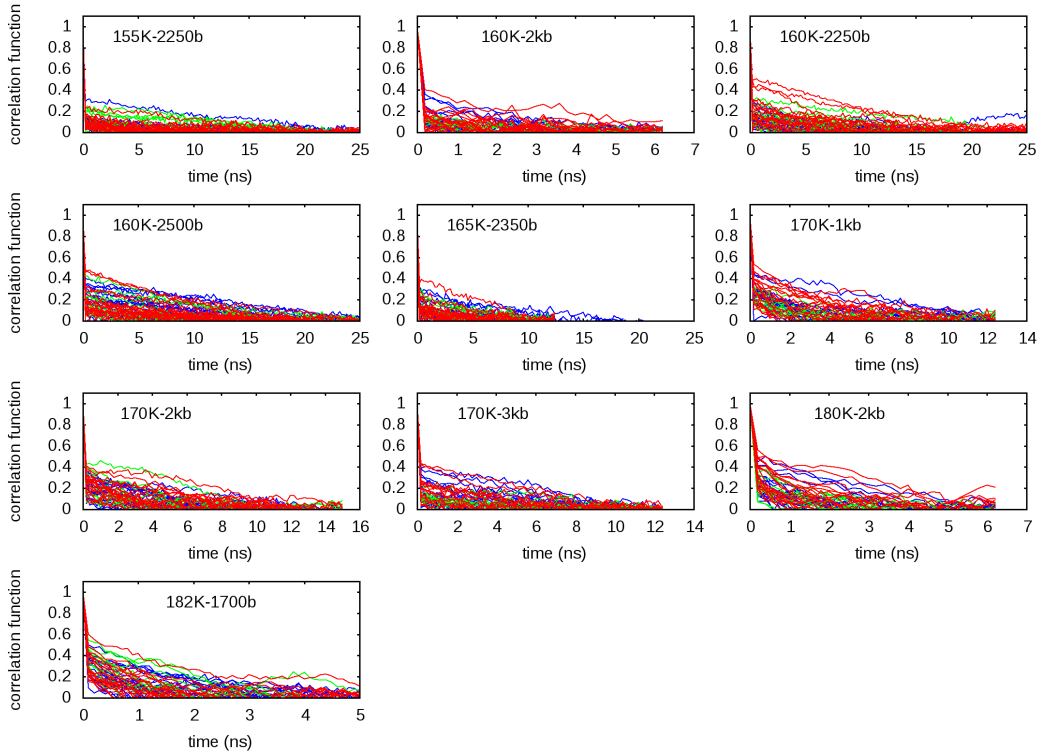


FIGURE 6.3.4 – Normalized self-correlation function $C_i(\rho(t))$ of the density in umbrella sampling simulations. Each line corresponds to an umbrella sampling window, with $S < 1.4$ in blue, $1.4 < S < 1.6$ in green, and $S > 1.6$ in red. The first fourth of each trajectory is not employed to compute the correlation, and the total length of each trajectory is four times the duration plotted.

Discarding the first half, we cut our trajectory into 2 to 10 blocks and computed the free-energy for each block. We used this to estimate the standard error for each block sub-division, and took the largest error among the different numbers of blocks. Those are represented as error bar on figure 6.3.2, with a typical size inferior to $1k_BT$. Concretely, their small size show that the free energy profile does not change much over the last half of the sampling time, which further assess the umbrella sampling convergence. Note that this is typically not true if you take first half of the trajectory, as free energy estimation can vary drastically during the relaxation of the S variable.

6.3.4 Schematic no man's land phase diagram

Figure 6.3.6.a summarizes these results in a schematic phase diagram that we reconstruct in no man's land, based on the structural and dynamical features of the low-free-energy part of configuration space, within 2 k_BT units from the minimum.

Average water densities and relative fluctuations at each P, T point are reported in figure 6.3.6.b and in figure 6.3.7a.

To distinguish liquid and amorphous phases, we followed a previous study on TIP4P/2005 water [136], based on self-diffusion coefficient D values and their decreasing trend. When entering the amorphous domain near 140 K, a one order of magnitude drop is expected for the diffusion, with $D \sim 10^{-14}$ cm^2/s [53]. Figure 6.3.7b shows the self-diffusion coefficient of each umbrella sampling windows, which decrease

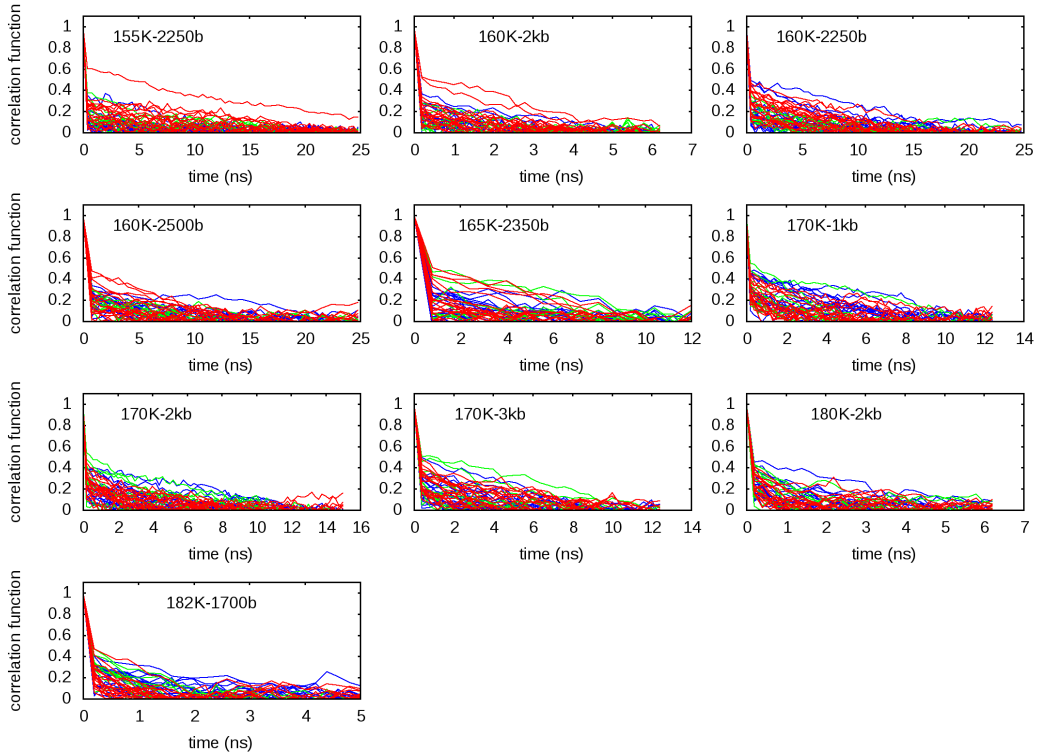


FIGURE 6.3.5 – Normalized self-correlation function $C_i(\langle Q_6(t)\rangle)$ of sixth-order Steinhardt parameter computed for a shell radius of 0.35 nm and averaged over all water molecules in umbrella sampling simulations [82]. Each line corresponds to an umbrella sampling window, with $S < 1.4$ in blue, $1.4 < S < 1.6$ in green, and $S > 1.6$ in red. The first fourth of each trajectory is not employed to compute the correlation, and the total length of each trajectory is four times the duration plotted.

with the temperature and drop of one magnitude over 20 K, while still being way above the low value of amorphous phases. Note that in the case of liquid transforming into amorphous, the magnitude drop will occurs on the same temperature. The two decreasing cause (thermal cooling or phase transformation) should not be confused. Also note that precise estimate of the relative phase stability and where the transition occurs would require computation of free energy profile.

6.4 Kinetic properties of the explored (P, T) conditions

At this point, the natural question becomes : do the large density fluctuations in the white band of Fig. 6.3.6 correspond to coexistence of two distinct water forms, low-density and high-density, and hence two metastable states ? To address this relevant issue we generated tens of long *free and unbiased* molecular dynamics trajectories. Starting from selected umbrella sampling configurations of type low-density and high-density liquid, to observe the spontaneous relaxation of the system and the coherence with respect to umbrella sampling free-energy landscapes. We generated the following trajectories, with a cumulative duration of more than 145 microseconds : 15×5,000 ns at 160 K, 2.5 kbar ; 7×4,000 ns at 160 K, 2.25 kbar ; 10×500 ns at 170 K, 1 kbar ; 10×2,000 ns at 170 K, 2 kbar ; 10×500 ns at 170 K, 3 kbar ; 10×1,500 ns at 180 K, 2 kbar.

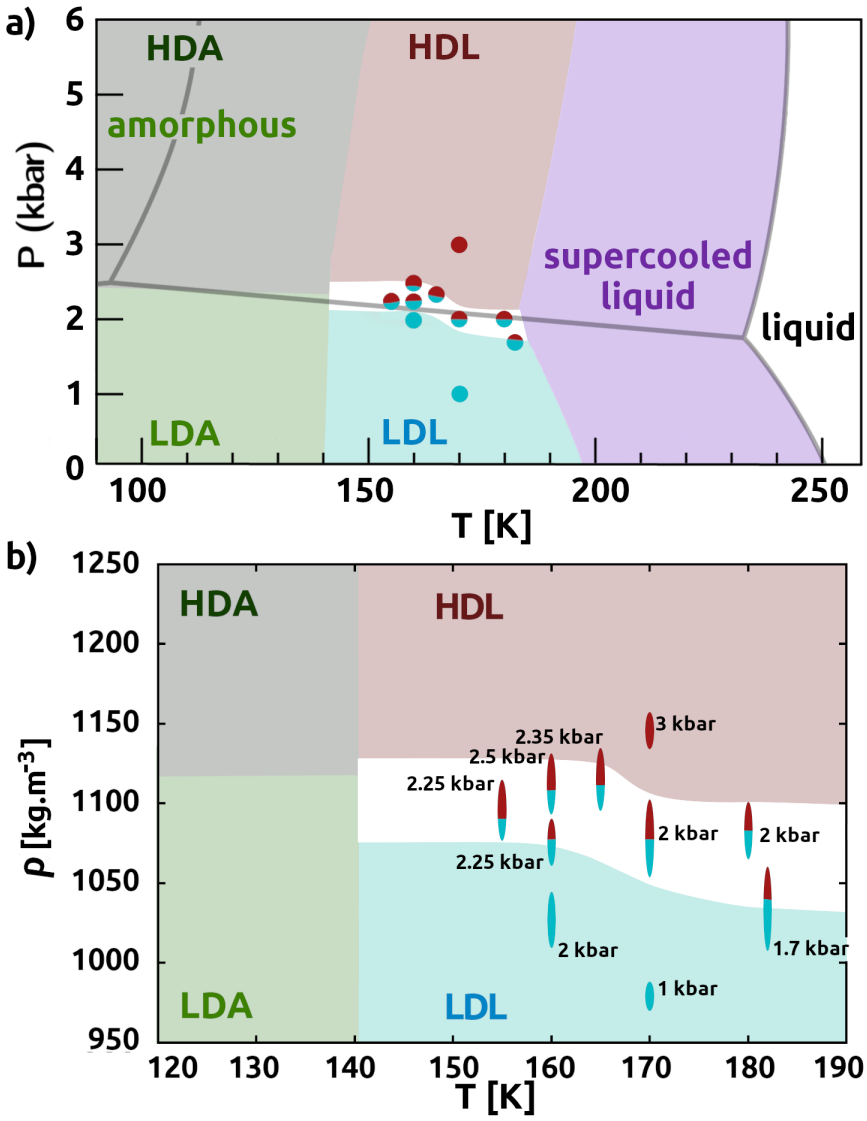
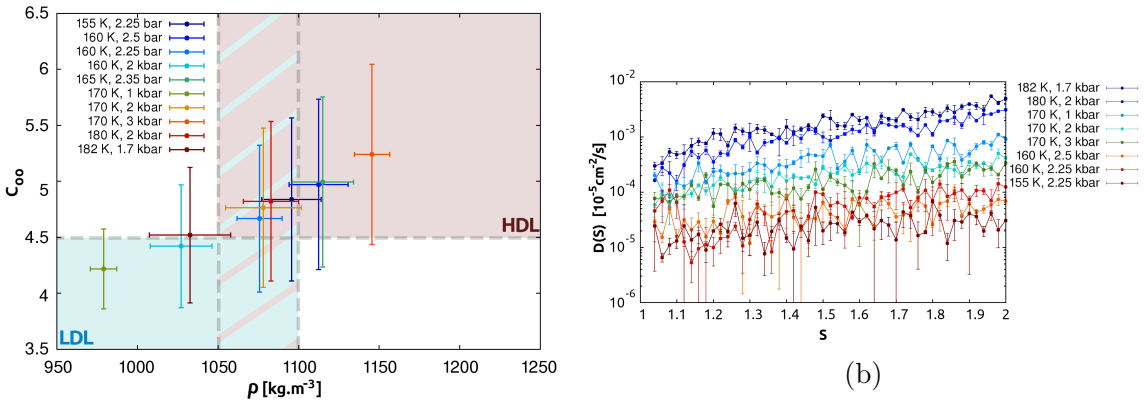


FIGURE 6.3.6 – (a) Schematic phase diagram of TIP4P/2005 water in the P, T region considered in this work. The solid gray lines indicates the stable phases [74]. Dots represent the conditions of MD simulations. Blue and red dots correspond to LDL and HDL, respectively. Dots half-red and half-blue indicate a nearly flat free-energy profile spanning low- to high-density water. The same color scheme is adopted to indicate areas where each of the two forms is expected to prevail. In the white areas the system is neither clearly LDL nor HDL. Low- and high-density amorphous forms are indicated in light- and dark-green colors, respectively. (b) Density as a function of temperature for the same phase diagram. The average density and its standard deviation (height of the ellipsoids) are computed by re-weighting the density values in umbrella sampling simulations with the equilibrium population $\mathrm{e}^{-G(S)/k_\mathrm{B}T}$, as a functon of the S path coordinate.

Figure 6.4.1 shows unbiased trajectories at 170 K initiated from the end-point of low or high-density umbrella sampling simulations. Comparison with Figure 6.3.2.a demonstrates that MD trajectories behave as expected from the computed free-energy profiles, relaxing from high- towards low free-energy regions according to the slope



(a) (b)

FIGURE 6.3.7 – (a) Average oxygen-oxygen coordination number C_{OO} (see section 6.5.1) as a function of the average density ρ for the various (P, T) condition explored in this study, computed from the weighted contribution of every umbrella sampling window with the Boltzmann factor $\mathrm{e}^{-G(S)/k_{\mathrm{B}}T}$ to obtain equilibrium distribution. The respective fluctuations are shown as horizontal and vertical bars (standard deviation of the distributions). The horizontal dashed line indicate the criterion used to separate low and high density liquid, according to coordination number [81]. The two vertical dashed lines indicate the extreme values of the density at $S = 1.5$ as shown in SI Fig. 2, as a criterion to separate LDL and HDL regions. (b) Diffusion coefficients of oxygen atoms for several P, T conditions, computed from the mean square displacement with gromacs on the biased umbrella sampling trajectories for each windows. The first half of the trajectory is discarded as equilibration. The error bar represent the standard deviation.

of the profile (i.e., the mean force), until showing stationary free diffusion in the region of the minimum. The latter is well-localized at low density at 1 kbar, it has a broad shape at 2 kbar, and is well-localized at high density at 3 kbar, as discussed above. As a further quantitative benchmark, the density distributions reconstructed from unbiased trajectories are in good agreement with those reconstructed from the equilibrium free energy profiles obtained by umbrella sampling (see Fig. 6.4.2).

Hence, unbiased MD is consistent with enhanced sampling simulations and it represents an independent robust validation of the reconstructed free-energy landscapes. Once again, we never observe local kinetic trapping of the system in two distinct states : at all P, T conditions and irrespective of the starting density the system steadily relaxes towards a single precise region in configuration space, without evident bottlenecks.

Note however that for low temperature ($< 165K$) our statistics became relatively limited, and that it would be desirable to lengthen the unbiased trajectories. More specifically, in Fig. 6.4.1.e and f, diffusion within the $2k_{\mathrm{B}}T$ free-energy minimal area are really sluggish, so that we don't observe large oscillation of the trajectories in the limited duration of our simulation (contrary to what could be observed in Fig. 6.4.1.c and d). Furthermore, its important to start from well equilibrated low or high-density state, as one could argue that only a truly well equilibrated states is metastable. This come back to assess the convergence of our umbrella samplings, as we take their

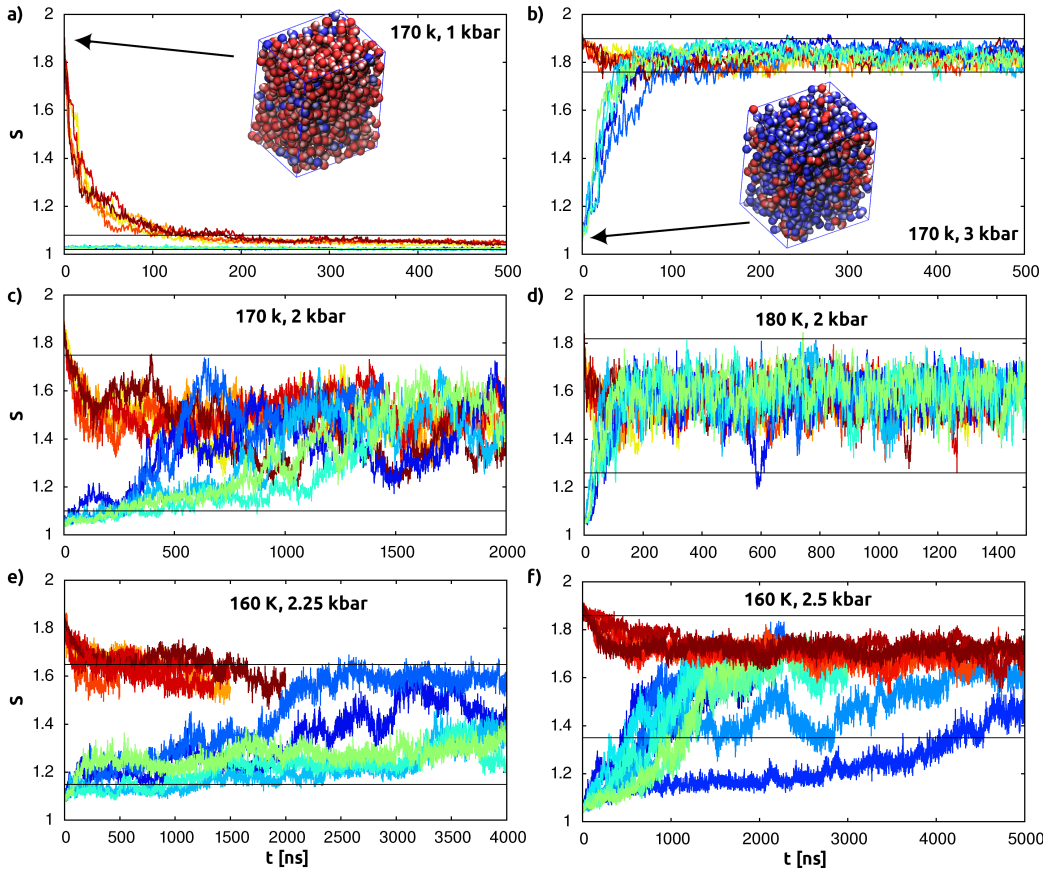


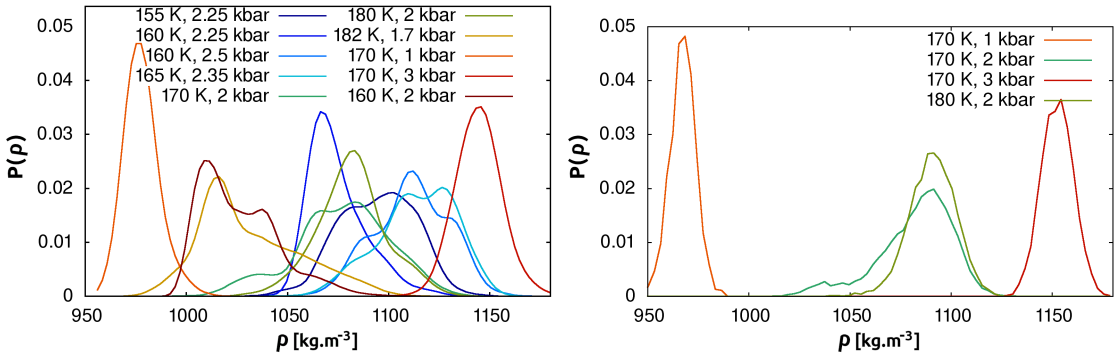
FIGURE 6.4.1 – Independent unbiased MD trajectories initiated from LDL (blue) or HDL (red) umbrella sampling end-point configurations, at 170 K and three different pressures (note the different horizontal scales). The intervals delimited by black lines correspond to free-energy values within 2 $k_{\mathrm{B}}T$ from the minimum as reconstructed from umbrella sampling (figure 6.3.2.a), where all unbiased trajectories converge, regardless of their initial configurations. (a, b) for well defined minimum, configuration relaxes within $\approx$ 100 ns. (c, d, e, f) for flat free energy profile with broad minimum, relaxaiton depend on the temperature. For 180 K at 2 kbar, relaxaion occurs within 100 ns. For 170 K at 2 kbar, it occurs within 2μs. For 160 K at 2.25 or 2.5 kbar, it occurs on the μs scale.

end-states for the initial configurations of our unbiased trajectories. Even if we are confident in it as discussed in section (6.3.3), it would not hurt to lengthen them for $T < 165K$.

6.5 Structural properties

6.5.1 Coordination number

We computed the oxygen-oxygen coordination number C_{OO} as the number of neighbors within a cutoff of 0.34 nm, using PLUMED with the following switching function : $\mathsf{c}(r) = (1-(r/0.34)^{32})/(1-(r/0.34)^{64})$. Next, we time-averaged the coordination number for each atom over time intervals of 20 ps along the umbrella sampling trajectory and computed the probability distribution. We tested several time interval



(a) (b)

FIGURE 6.4.2 – Distribution of density ρ for various (P, T) condition explored in this study. (a) Computed as described for the C_{OO} distribution from umbrella sampling in section (6.5.1), except that we do not perform any time average here. (b) Estimated from unbiased simulations presented in figure 6.4.1. The first half of the trajectories is discarded as equilibration. The distributions are consistent with those computed in (a).

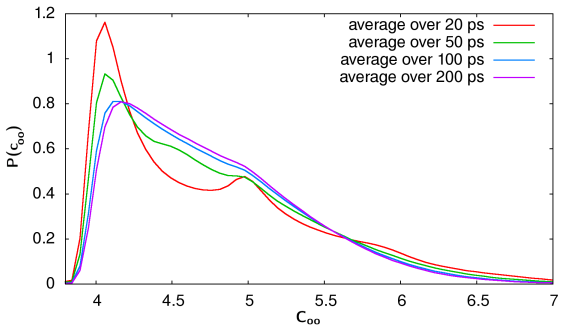


FIGURE 6.5.1 – Coordination number probability at 170 K and 2 kbar from umbrella sampling simulations averaging C_{OO} of each oxygen atom over four different time intervals.

as reported in figure 6.5.1, which shows that even if the moment of the distribution are not affected by this averaging, it shape is, with disappearance of the slightly bimodal nature of the distribution. In this way we built one histogram for each umbrella sampling window, which was smoothed to reduce irrelevant noise, averaging over the adjacent bins. To obtain equilibrium populations, we summed the re-weighted contribution of each umbrella sampling window according to the Boltzmann factor $Z^{-1}\mathrm{e}^{-G(S)/k_\mathrm{B}T}$.

Figure 6.5.2 presents the resulting distribution of C_{OO} for all the P, T conditions explored in this study. For those that have a well localized free energy minimum, see Fig. 6.3.2.a, distributions have one well defined peak when low-density liquid is favored, and a broad almost flat distribution when high density is favored. Whereas those that present relatively flat free energy profile, see Fig. 6.3.2.b, distributions are broad and slightly bimodal, presenting features of both low and high-density liquid. Its important to note that one of the initial argument to claim the existence of a first-order liquid-liquid transition were based on such bimodality [81]. Here we see that slight bimodality does not necessarily imply the existence of an underlying free

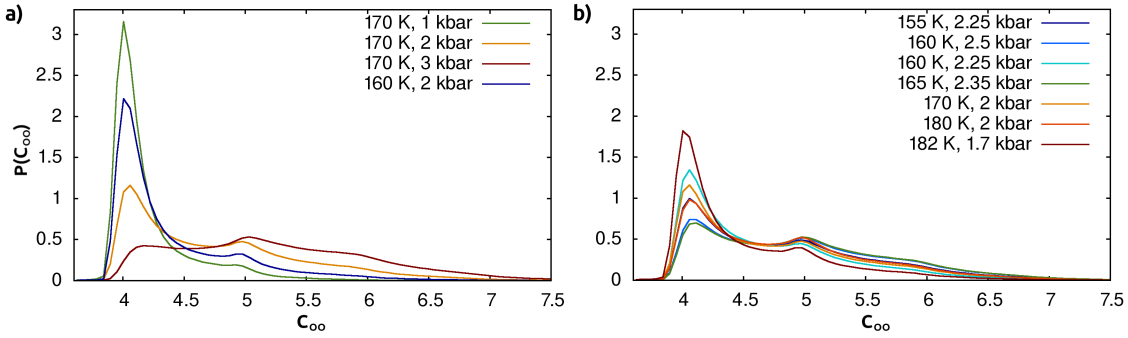


FIGURE 6.5.2 – Distributions of oxygen-oxygen coordination numbers C_{OO} (a) Pressures and temperature around 170 K and 2 kbar. (b) Conditions intermediate between those favoring low density and high density.

energy barrier.

6.5.2 Cluster analysis

In Figure 6.5.3 and 6.5.4 we traced the number of bulk molecules in a spherical drop of LDL or HDL water by counting, in the reference structures at 170 K and 0 or 5 kbar, how many molecules belonging to a sphere (mathematically defined within a bulk periodic configuration) are in contact only with molecules of the sphere itself and not with external molecules. Contact is defined using the same switching function $\mathsf{c}(r)$ discussed above. On the opposite side, as a reference limit for the case of random mixing between LDL and HDL molecules we generated random bond networks with the same distribution of coordination numbers as obtained from MD simulations, starting from a random initial adjacency matrix and adding/removing random bonds (10^5 Metropolis Monte Carlo steps) until a deviation

$$\int dC_{OO} |P_{MD}(C_{OO}) - P_{RN}(C_{OO})| = 0.058 \pm 0.007 \tag{6.5.1}$$

between the probability distributions of coordination numbers from MD and from the random network.

As a final benchmark, we analyzed the structure of instantaneous atomic configurations, with particular attention to P, T conditions maximizing density fluctuations, to understand whether low- and high-coordinated water molecules are randomly mixed or they group together in order to minimize the LDL/HDL interface. Clearly, the hypothesis of a coexistence of two *distinct* liquid forms requires the existence of a well-defined geometrical interface characterized by unfavorable molecular interactions, hence of minimal extension (spherical or planar). Under such hypothesis, as in classical nucleation theory, the interface provides an unfavorable free-energy contribution to the total budget of the system, creating a barrier that grows with system size as $N^{2/3}$ (as observed in ST2 water in Ref. [23]).

Visual inspection both of unbiased MD trajectories and of umbrella sampling trajectories does not reveal a clear tendency towards separation of large and convex LDL or HDL regions : the respective clusters of hydrogen-bonded molecules display a complex, interpenetrating interface whose extension appears far from minimal (see Fig. 6.5.5). A quantitative assessment is presented in Figure 6.5.3 and 6.5.4 : molecules

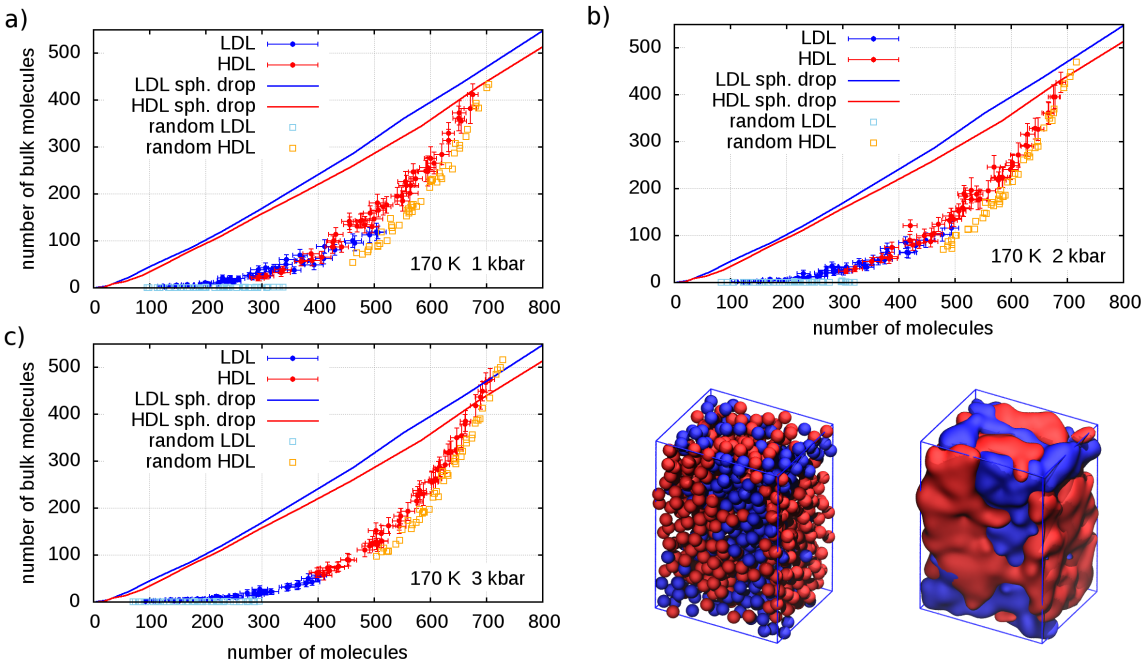


FIGURE 6.5.3 – Number of LDL-like and HDL-like molecules (assigned on the basis of $C_{OO} < 4.5$ or > 4.5, respectively) that are surrounded by molecules of the same type, *i.e.* not at the interface LDL/HDL, extracted from umbrella sampling trajectories. For comparison, continuous curves indicates the number of bulk molecules in a spherical droplet containing only LDL or HDL, and the squares correspond to random networks of molecules with the same bond distribution as LDL or HDL in MD configurations (see section Materials and Methods for details). The 3D structure (balls and surfaces enclosing them) illustrate a typical configuration at 170 K and 2 kbar.

are identified as LDL-like or HDL-like based on $C_{OO} < 4.5$ or > 4.5, respectively, and for each type the number of bulk molecules (i.e., in contact only with alike molecules, thus not at the interface) is plotted against the total number. In principle, the fraction of bulk molecules is maximized when all molecules of one type form a single spherical drop (or a flat periodic slab), and it is minimized when molecules are randomly mixed. These two limits are also represented in Figure 6.5.3 and 6.5.4, allowing to appreciate how MD configurations at putative coexistence conditions (i.e., with similar LDL and HDL fraction) are in reality much closer to a randomly intermixed system than to one exhibiting phase-separation. We find similar results at all P, T conditions explored, both for umbrella sampling and unbiased trajectories, whenever both LDL and HDL are present in significant amount. Previous studies addressed the number and size of LDL/HDL clusters in TIP4P/2005 water, albeit at $T \geq 190$ K and without discussing the interface shape [139]. In summary, our structural analysis is once again consistent with the absence of liquid-liquid phase separation.

6.6 Discussion

We performed enhanced sampling and unbiased MD simulations in a range of P, T conditions between 155 and 182 K and between 1 and 3 kbar, and in all cases we could not find any compelling evidence of liquid-liquid phase separation and of a

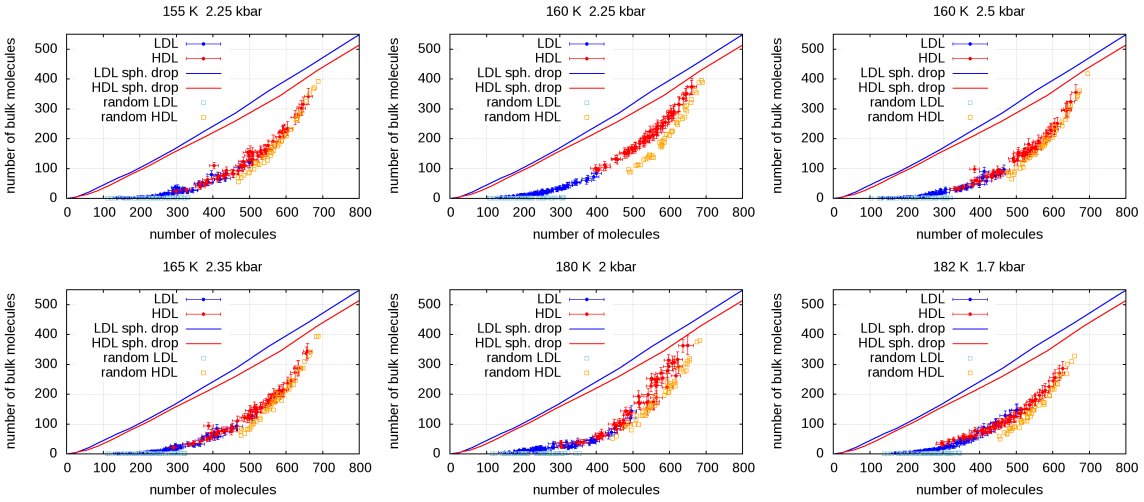


FIGURE 6.5.4 – Number of LDL-like or HDL-like molecules (assigned on the basis of $C_{OO} < 4.5$ or > 4.5, respectively) that are surrounded by molecules of the same type, i.e. not at the interface LDL/HDL, within umbrella sampling trajectories. For comparison, continuous curves indicates the number of bulk molecules in a spherical droplet containing only alike molecules, and the squares correspond to random networks of LDL-like or HDL-like molecules with the same bond distribution as MD configurations (see Materials and Methods for details).

corresponding free-energy barrier. We reach this conclusion employing three different and complementary methods : 1) enhanced sampling simulations to reconstruct free-energy landscapes for the low- to high-density transition, 2) long unbiased MD simulations to probe the putative local stability of LDL and HDL phases and to confirm free-energy landscapes, and 3) in-depth structural analysis of clusters formed by low- and high-density water to assess the geometric properties of the LDL/HDL interface, a crucial indicator of phase separation.

In particular, at P, T conditions close to the most recently predicted locations of the liquid-liquid critical point (182 K, 1.7 kbar and 180 K, 2 kbar to compare with Ref. [24], and 170 K, 2 kbar to compare with Ref. [26]), we found no free-energy barrier and a single broad minimum (Fig. 6.3.2.c), characterized by significant fluctuations in density and coordination number (Fig. 6.3.6.b, 6.3.2.d), without evidence of phase separation between LDL and HDL. These qualitative features, however, are not unique of a single point in P, T-space, since we could follow a line of points with similar behavior – in particular without free-energy barrier – from 182 K down to at least 155 K, close to the frontier with amorphous water.

In the recent study Ref. [26], MD simulations with the TIP4P/2005 and TIP4P/ice potentials are performed at supposed supercritical conditions (above 177K for TIP4P/2005) to infer the existence of critical behavior and of a critical point at lower temperature (about 172K and 1.9 kbar), by temperature extrapolation with an a-posteriori reweighting procedure and by comparison with an idealized model (3D Ising). The scope of the latter work is quite different from ours, where we directly probe (without extrapolation) the low-temperature behavior of water, down to 155 K, by means of both unbiased MD and enhanced sampling, drawing factual observations about our results without making hypotheses based on a model. Even if the conclu-

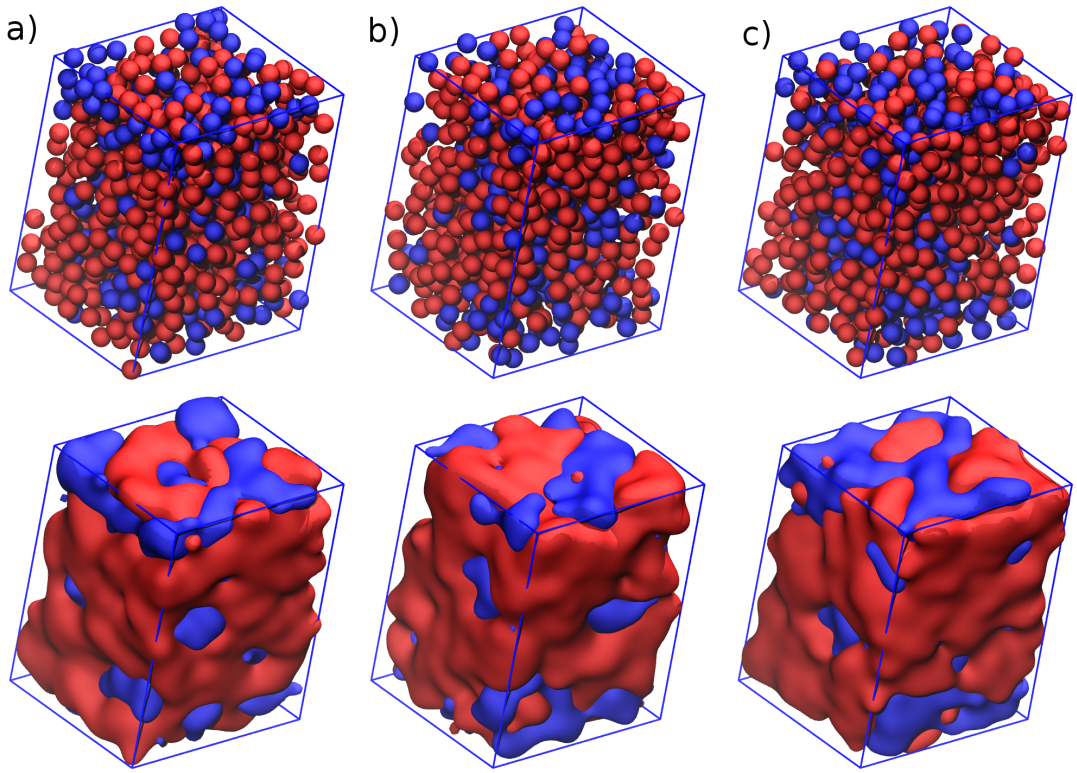


FIGURE 6.5.5 – Example of 3D structures from unbiased MD at 160 K, 2.25 kbar. LDL-like molecules (with $C_{OO} < 4.5$) are indicated as blue balls, HDL-like ones (with $C_{OO} > 4.5$) as red balls ; surfaces are drawn with the "QuickSurf" tool of vmd (isosurfaces extracted from a volumetric Gaussian density maps on a uniformly-spaced 3-D lattice) [137, 138].

sions of the Authors of Ref. [26] (existence of a critical point and of a liquid-liquid phase transition) are different from our own observations (lack of free-energy barrier), still the physical quantities directly computed in both works display good agreement : this is the case for instance if we compare the equilibrium distribution of densities at 178K, 1.7 kbar and 2 kbar in Fig. S1 of Ref. [26] with those at 182K, 1.7 kbar and at 180K, 2 kbar in Fig. 6.4.2 of our work. Instead, the bimodal distributions in Fig. 1B and 2A in Ref. [26], at lower temperature down to 171 K, are extrapolated based on a reweighting technique. The distributions are still compatible with our results : despite a visually clear bimodal shape, the probability minimum between the two peaks is not deep, so that conversion to free energy profiles as $F(\rho) = -k_{\mathrm{B}}T \log(P(\rho))$ imply barriers smaller than 1.2 $k_{\mathrm{B}}T$ (what can be hardly defined as a barrier at all). This is coherent with the approximately flat free energy profile and its error bars presented in our Fig. 6.3.2 for 170 K at 2 kbar, within about 1 $k_{\mathrm{B}}T$. We also note that the density fluctuations displayed in Fig. 1 of Ref. [26] are obtained with a box size of only 300 molecules. Clearly, if the latter point in the phase diagram was close to a critical point, a barrier of increasing height should be observed for decreasing T, however our simulations down to 155 K do not show any sizable barrier.

The situation is different for the ST2 model : in Ref. [23] a 4 $k_{\mathrm{B}}T$ free-energy barrier separating LDL from HDL could be measured at 229 K and 2.4 kbar for a system size of 192 molecules, based on extensive and careful enhanced sampling simulations ;

the barrier was confirmed by unbiased Monte Carlo simulations reversibly sampling the LDL-HDL transition, and it was shown to scale like $N^{2/3}$ for $198 \leq N \leq 600$, as expected for a first-order phase transition. We remark that we could not find any barrier with the more accurate TIP4P/2005 force field despite a system size of 800 molecules, larger than the largest one considered in Ref. [23].

From the viewpoint of the physics of supercooled water, the most important lesson delivered by freely relaxing trajectories is that there is no P, T point (within the broad range we explored) where LDL and HDL are both kinetically trapped in their respective forms for a measurable time (Fig. 6.4.1). We must conclude that it is impossible to observe LDL and HDL as distinct and persistent forms at the same thermodynamic conditions. In other words, coexistence of the two phases is impossible, so that LDL and HDL are not two distinct phases in the thermodynamic sense. We remark that while the existence of a mechanically stable LDL/HDL interface has been demonstrated for the ST2 model up to large system sizes [131, 140], such a demonstration is lacking for the more reliable and accurate TIP4P force fields family [1].

On the contrary, we conclude that it is possible to change the form of water from lower-density and lower-coordination values to higher ones in a continuous way, for instance by increasing pressure from ≈1.5 to ≈3 kbar at any temperature between 155 K and 182 K (Fig. 6.3.2), without encountering a bottleneck in phase space, i.e., a barrier. Of course the timescale necessary for the system to relax from an initial out-of-equilibrium density slows down when lowering T, from ≈500 ns at 170 K and $1 - 3$ kbar to ≈2 μs at 160 K and $2.25 - 2.5$ kbar (see Fig. 6.4.1), however such a slow evolution appears the result of continuous diffusion in density space with a weak diffusion coefficient, rather than of Poisson-distributed rare jumps across a barrier. This factual observation of the behavior of unbiased MD trajectories is fully consistent with our enhanced sampling simulations, where no free-energy barrier could be measured, and also with our analysis of the three-dimensional structure of low- and high-density water clusters, that revealed no strong tendency to minimize the interface area and a situation closer to random intermixing of LDL-like and HDL-like molecules than to phase separation (Fig. 6.5.3 and 6.5.4).

6.7 Conclusions and outlook

Our results show that the metric based on permutation invariant vectors [7, 8] resolves well the range of supercooled water structures and densities throughout the vast P, T region explored. This result extends the analyses in Ref. [84], where the same metric was demonstrated able to resolve and clusterize structures belonging to liquid, amorphous and crystalline water. In combination with path coordinates, the metric allowed here also to reconstruct free energy landscapes extending the approach of Ref. [8], applied also to heterogeneous ice nucleation in Ref. [90], to transitions between supercooled liquid forms. While several other order parameters have been applied to specific investigations on water [111, 76, 40], due to its generality our computational approach allowed a comprehensive and unitary study of water structure, dynamics and thermodynamics encompassing liquid polymorphs, solid polyamorphs and crystals.

We do not observe the kinetic trapping (hence metastability) of the LDL and

HDL forms at the same thermodynamic conditions, hence coexistence of two phases. This observation is not compatible – at least for the system size we considered – with hypotheses evoked in the literature on the existence of an unfavorable interfacial free energy preventing the formation of two liquid phases in finite-sized systems, or on phase-separation dynamics much slower than simulation times of the order of hundreds of nanoseconds [131, 57, 24]. All our results indicate the lack of a first-order liquid-liquid phase transition and of the related critical point for the accurate TIP4P/2005 water force field, thus leading to discard the multiple scenarios that include such features and that have been hypothesized in the last 40 years to explain water anomalies [111, 1]. Notwithstanding the unquestionable importance of simplified theoretical models to help us understanding complex phenomena, this conclusion underlines the difficulty in extrapolating observations from other regions of the phase diagram deep into no man's land.

However, future studies could take advantage of the growing computing power to accumulate more extensive numerical data on supercooled water and improve the overall statistics. Even if all our analyses (correlation functions, block averaging, comparison with unbiased trajectories) indicate that the free energy landscapes we report are converged, it would be interesting to extend the duration of umbrella sampling windows to the microsecond time scale, especially in the deeply supercooled region ($T \leq 160K$), where relaxation of the system became really sluggish. Similarly, longer unbiased trajectories would display more clearly the diffusion properties of the order parameter in the latter thermodynamic region. We also remark that our study addressed only one – albeit relatively large – system size (800 molecules), due to the need to explore an extended P, T region in no man's land, while in future studies it would be interesting to extend similar simulations to larger sizes.

Finally, an important aspect that remains to be elucidated is the evolution of the free-energy landscape upon cooling below 140 K, where the LDL/HDL transition becomes the LDA/HDA transition between amorphous forms, with previous experiments and simulations indicating first-order character and the existence of a barrier. Rigorous free-energy calculations would be even more computationally demanding than for supercooled liquid water, but today they are becoming accessible. It appears that due to the rich and peculiar phenomenology of water physics and chemistry, the inescapable primary sources of information remain today experiments and atom-detailed computer simulations that directly probe the P, T conditions of interest.

7 Study of the Homogeneous Ice Nucleation

7.1 Introduction

Among the many stable and metastable solid states in which water can be, the ice I polymorphs (hexagonal, cubic and stacking disordered) are those of most interest for us, as they are the one that occurs in atmospheric temperature and pressure conditions on earth. So even if water can nucleate into several ice forms as shown in figure 7.1.1, when one speaks about ice nucleation it generally means the nucleation of ice I. Here we are mainly interested to the homogeneous ice nucleation problem, that is with only pure water, free from any kind of impurities or interfaces which could serve as nucleation site to speed up the process.

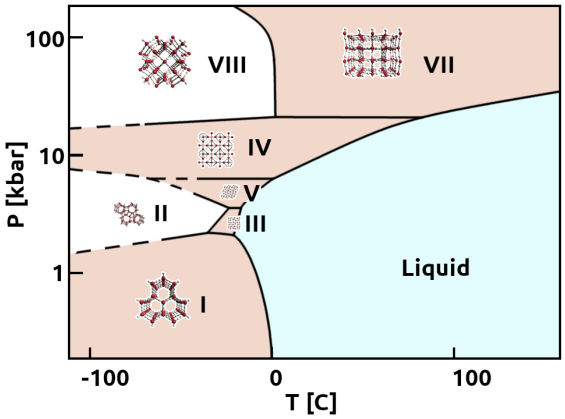


FIGURE 7.1.1 – Partial stable phase diagram of water, where we have represented in orange the crystalline ice in which liquid water can nucleate.

The most widespread theory that exist to describe nucleation is the classical nucleation theory (CNT), so even if we did not make a direct use of it, we will present it briefly as it sheds light on the general processes occurring in nucleation. Then we will present the ice I polymorphs and their crystalline properties, as this knowledge is required to properly understand the homogeneous nucleation process in water.

7.1.1 Classical nucleation theory

Classical nucleation theory, even if partially, is referred to in almost every computer simulation. Here we are just going to sketch the broad lines, while a thorough description of this theory has been made elsewhere for the willing reader [141, 142].

This theory was first created to describe the condensation of supersaturated vapors into liquid, but its concepts can be applied to nucleation of crystals from supercooled liquids. Following classical nucleation theory, clusters of crystalline particles (either atoms or molecules) of any size are treated as large homogeneous pieces of crystalline phase, with a thin interface with the surrounding liquid. This hypothesis is known as the capillarity approximation and it concretely means that we are neglecting any effect due to the detailed atomic structure of the cluster, including crystal lattice structure, as well as the possible role of different polymorphs

or structural defects, particularly in the interface region. Hence under this strong assumption, the nucleation process can be fully described by the interplay between interfacial free energy σ estimated in the approximation of an infinite planar interface and the free energy difference between the liquid and crystalline phase per unit volume Δg_v. When nucleation occurs in three dimension, the free energy cost ΔG to form a spherical crystalline nucleus of radius r is therefore the sum of a surface and a volume term

$$\Delta G = \underbrace{4\pi r^2 \sigma}_{\text{surface}} - \underbrace{\frac{4\pi r^3}{3} \Delta g_v}_{\text{volume}} \tag{7.1.1}$$

where the volume of the crystalline nucleus V is by definition equal to the number of crystalline particles N divided by the solid density ρ_s, $V = (4/3)\pi r^3 = N/\rho_s$. This function has a maximum at the critical nucleus size N_c

$$N_c = \frac{32\pi\rho_s}{3} \left(\frac{\sigma}{\Delta g_v} \right)^3 \tag{7.1.2}$$

N_c is the number of particles that a crystalline cluster must include to overcome the cost due to the formation of a solid-liquid interface. Before complete crystallization of the system, small crystalline clusters will occur due to infrequent spontaneous fluctuations. Eventually a sufficiently large cluster will form, overcoming the free energy barrier for nucleation, that we can compute by inserting the expression of N_c into the definition of ΔG

$$\Delta G_c = \frac{16\pi}{3} \left(\frac{\sigma^3}{\Delta g_v^2} \right). \tag{7.1.3}$$

One of the assumptions of this theory is that once a crystalline nucleus has reached this critical size, it will extend itself to the whole system.

To describe the kinetics of nucleation, classical nucleation theory further assumes that no correlation exists between successive growing or shrinking events of a nucleus. Concretely, this means that the evolution of the nucleus is considered as a Markovian process, where particles attach or detach themselves from the crystalline cluster independently. This also suppose that the crystal does not undergo important structural change during the typical duration that a nucleus take to extend itself to the whole system. Furthermore it requires that thermal history of the system plays no significant role in the nucleation process, which solely depends on the temperature and pressure [143, 144]. Under all of these assumptions, we can compute the homogeneous nucleation rate per unit volume, that is the probability by unit volume to form a critical nucleus, with the following expression

$$\mathcal{J} = \rho_l \mathcal{A} \mathcal{Z} \exp\left(-\frac{\Delta G_c}{k_\mathrm{B} T} \right) \tag{7.1.4}$$

where ρ_l is the liquid phase density, which effectively represent the number of possible nucleation sites per unit volume. $\mathcal{A}$ is the attachment rate at which particles attach to the growing nucleus and is related to the time τ_λ required for a molecule to diffuse over a given length λ

$$\mathcal{A} = \frac{4N_c^{2/3}}{\tau_\lambda} \tag{7.1.5}$$

where we can compute τ_λ from the self-diffusion coefficient D as $\tau_\lambda = \lambda^2/6D$ [33, 35]. $\mathcal{Z}$ is the Zeldovich factor, which is there to account for the fact that a post-critical nucleus (with size $N > N_c$) might still shrink without growing into crystalline phase. It is related to the curvature of the free energy barrier

$$\mathcal{Z} = \sqrt{\frac{|\Delta g_v|}{(6\pi k_\mathrm{B} T N_c)}} \tag{7.1.6}$$

Classical nucleation theory allow us to link and calculate the two thermodynamic and kinetic quantities of highest interest, namely the free energy barrier ΔG_c and the nucleation rate $\mathcal{J}$. For the evaluation of ΔG_c, one can use rigorous direct free energy calculation like umbrella sampling or metadynamics, or use indirect approach like seeding. In the latter case, results rely heavily on the correctness of equation 7.1.2 and 7.1.3. The idea is as follows : we evaluate critical nuclei size as described in section (5.2.1), then we use thermodynamic integration to compute Δg_v. From there we can compute the interfacial free energy σ with 7.1.2, as ρ_s is easily evaluated with short simulation in NPT ensemble. Then all the pieces are put together to compute the free energy barrier using equation 7.1.3. Once the free energy barrier is estimated, all the other quantities required to compute the nucleation rate according to equation 7.1.4 can be easily evaluated with short simulations [33, 34, 36]. It is important to note that all these quantities, N_c, Δg_v, ρ_l, ρ_s, $\mathcal{A}$ are dependent of the temperature.

We want to stress again that all the assumptions made for classical nucleation theory break in a numerous number of systems [144]. As we will discuss the properties of Ice I and how homogeneous nucleation occurs in nature for water in the next sections, it will become clear that this theory is not the most well fitted to study nucleation of water, as most of its assumptions fail for this specific material.

7.1.2 The ice I polymorphs

Ice I has two well defined states : stable hexagonal ice ($\mathrm{I_h}$) and metastable cubic ice ($\mathrm{I_c}$). Until quite recently, observations of pure $\mathrm{I_c}$ has not been achieved and it was not clear if it could be observed in nature [28, 27]. Prior to this two studies, what was found in nature was a metastable stacking disordered phase, where hexagonal and cubic ice stack on of each other randomly. As this phase was first believed to be cubic ice until experiments proved that it was made of disordered stacks, it is sometimes called "ice $\mathrm{I_c}$" (with the quotes), but nowadays it is more commonly called stacking disordered ice $\mathrm{I_{sd}}$ [29, 30, 31]. Precise characterization of the disorder and its formation mechanism are hard to achieve experimentally, thus numerical studies are of prime importance to understand this phenomenon, and it is the aim of this work.

Both ice $\mathrm{I_h}$ and $\mathrm{I_c}$ are made up of identical layers of puckered six-membered rings of oxygen atoms connected with hydrogen bonds. It is the way these layers stack that distinguishes the two. As shown in figure 7.1.2, in ice $\mathrm{I_h}$ each successive layer is the mirror image of the preceding one. In ice $\mathrm{I_c}$ each layer is identical to the precedent, but shifted by one-half of the diameter of a hexagonal ring. The ice $\mathrm{I_{sd}}$ is a superposition of hexagonal and cubic layers. All cubic faces are equivalent, but only the basal face of hexagonal ice can stack seamlessly with the cubic ice. These differences in stacking results in different crystalline materials [31].

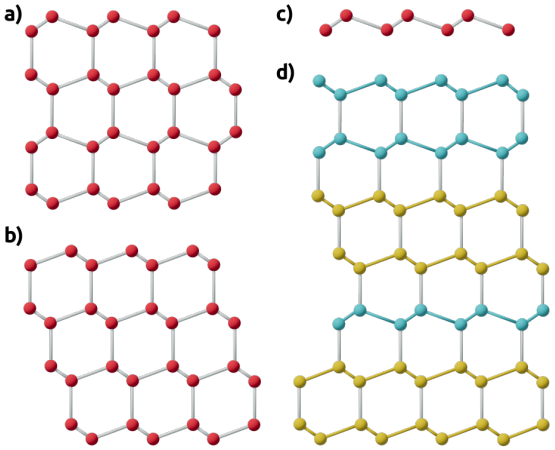


FIGURE 7.1.2 – Representation of the stacking of layers in ice I polymorphs. Spheres represent oxygen atoms and they are connected by hydrogen bonds. For clarity hydrogen atoms are omitted. (a) Crystal structures of hexagonal ice ; (b) crystal structure of cubic ice ; (c) a layer of ice ; (d) a possible arrangement of stacking disordered ice : hexagonal layers are colored in blue and cubic layers in yellow. Adapted from [31].

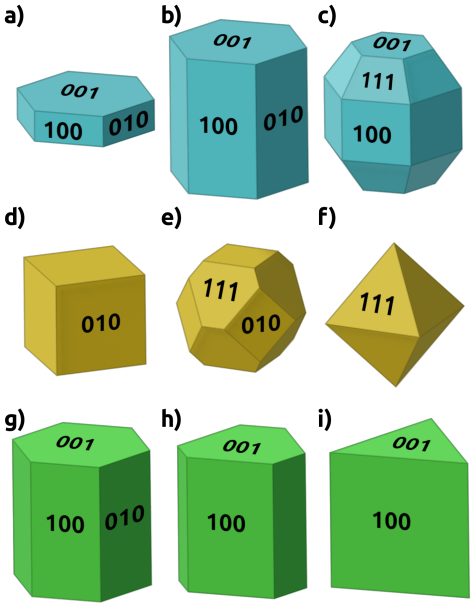


FIGURE 7.1.3 – Possible macroscopic crystal shapes for the ice I polymorphs depending on which faces grow the fastest. (a-c) are hexagonal, (d-f) are cubic and (g-i) are stacking disordered. Some of the Miller indices have been indicated. Adapted from [145].

7.1.3 A bit of crystallography

A crystalline material is characterized by its symmetries. On a microscopical scale, these symmetries are encoded in a crystalline space group. Ice I_h space group is $P6_3/mmc$, meaning that it has a primitive unit cell (P), with a six-fold screw axis (a rotation around an axis coupled with a translation along the axis, noted 6_3), two mirror planes (a simple reflection along a plane, noted m) and one glide plane (a translation followed by a reflection along a plane, noted c). If we stripe off the translation, that is by converting screw axis into rotational axis and glide plane into mirror plane, we will obtain its point group $6/mmm$, which characterize the symmetry of a macroscopic crystal, and thus its shape. Ice I_c space group is $Fd\bar{3}m$, meaning that it has a face centered unit cell (F), with a glide plane (d), a three-fold rotoinversion axis (a rotation followed by a reflection along a plane perpendicular to the axis, noted $\bar{3}$) and a mirror plane (m). Its point group is $m\bar{3}m$. Ice I_{sd} space group is $P3m1$ and its point group is $3m1$ [145]. This information is summarized in table 7.1.1.

Ice	Space group	Point group	Crystal system
I_h	$P6_3/mmc$	$6/mmm$	hexagonal
I_c	$Fd\bar{3}m$	$m\bar{3}m$	cubic
I_{sd}	$P3m1$	$3m1$	trigonal

TABLE 7.1.1 – Crystallographic properties of the Ice I polymorphs

In fact the pure states are named following their crystalline symmetry, as I_h lies in the hexagonal symmetry group and I_c lies in the cubic symmetry group. I_{sd} lies in the trigonal symmetry group, but we prefer to name it to emphasize its stacking disordered nature. The possible corresponding macroscopic structure of these symmetry group are presented in figure 7.1.3 [145]. The exact shape of the crystal is dictated by its internal symmetry and by the rates of growth of each of its crystal faces. To distinguish the various faces, we use Miller indices, that consists of three number (hkl). Faces with high Miller indices are less energetically favorable and thus crystals tend to prefer shape in which faces with lower Miller indices are exposed. For I_h and I_{sd} the faces with lowest Miller indices are the basal ones : $\{(001), (00\bar{1})\}$, and the prismatic ones : $\{(100), (\bar{1}00), (010), (0\bar{1}0), (001), (00\bar{1})\}$. The cubic nature of I_c renders all planes in the same class equivalent, meaning that all directions are equivalent.

7.1.4 Numerical study of nucleation of ice I

In nature homogeneous ice nucleation mainly occurs at supercooled conditions, and it is thought as a two step process : first water freezes into metastable stacking disordered ice I_{sd} and then the crystal will anneal at warmer temperature into the stable hexagonal ice I_h state, through rearrangement of the crystal lattice [146, 31]. This annealing is quicker with warmer supercooled temperature [147, 148]. The initial relative proportion of I_c and I_h, i.e., the cubicity, in I_{sd} is correlated with the temperature, lower temperature meaning higher cubicity [149]. Thus to understand homogeneous ice nucleation of water we need to understand the mechanisms that render I_{sd} easier to form than I_h, despite being metastable. As this involves time and space resolutions far beyond those of experiments, numerical simulation seems better prepared to tackle this task. In this regard it is important to recall that results of numerical simulations are highly dependent of the choice of water model, as discussed in (3.2). Hence it is important to distinguish studies made with mW and those made with TIP4P/2005 or TIP4P/Ice, even though the models give coherent results.

In the last decades, due to the development of new enhanced sampling methods and of the mW model, which is at least 100 times faster than TIP4P-like models from a computational point of view, many works addressed homogeneous ice nucleation.

A first set of studies used brute force methods to estimate the free energy barrier and nucleation rate with mW models [150, 151]. In parallel, a large amount of work as been devoted to systematically study nucleation of water at different temperatures for several models (mW, TIP4P/2005, TIP4P/Ice, etc.), using the seeding technique to compute the free energy barrier, nucleation rate and the time to crystallize a complete system [32, 33, 34, 35, 36]. Even if TIP4P/Ice and mW give similar results for some thermodynamic quantities, they differ largely when looking at their kinetic properties. The growth rate is three orders of magnitude larger for mW than for TIP4P/Ice, and their nucleation rate differs by almost 10 orders of magnitudes, those of TIP4P/Ice being much closer to the experimental results [152, 36]. In practice all the works cited in this paragraph provide valuable knowledge for further studies, as they estimated the critical nuclei size and nucleation rate for a wide range of temperatures, allowing one to easily choose the right conditions of temperature and simulation box size to simulate nucleation.

Following the latter series of works, a series of studies based on more rigorous

enhanced sampling methods than seeding were published. Extensive forward flux sampling simulations were used to compute the nucleation rate and to study the prevalence of I_{sd} at 230 K for TIP4P/Ice, showing that presence of cubic ice leads to more compact crystallites. These are more likely to grow into bulk crystal system , compared to less compact and more hexagonal nuclei, and the critical cubicity was estimated to be $C = 0.59 \pm 0.07$. Even though, in principle, forward flux sampling should give a more correct estimation of the rate than brute force or seeding methods, as it doesn't rely on classical nucleation theory and can be directly computed from the sampled data using general mathematical properties of the stochastic process subjacent to forward flux sampling, the computed nucleation rate was off by more than 8 order of magnitude compared to experiments [153, 42]. Authors pointed out that it may be due to underestimation of the chemical potential difference between the liquid and hexagonal ice by TIP4P/Ice. Reproduction of this work with another variant of forward flux sampling revised this rate with 4 order of magnitude [37]. In fact despite their solid mathematical background, all forward flux sampling methods accuracy are dependent on their specific implementation and variants [154].

Using extensive aimless shooting path sampling simulations, it was shown that for small system size – up to 100,000 molecules – at 230 K stacking disordered ice is more stable for the mW model, with purely hexagonal nuclei evolving toward disordered ones during relaxation of the transition state ensemble. The critical cubicity was estimated to be $C = 0.63 \pm 0.05$, in agreement with forward flux sampling results on TIP4P/Ice [151, 38]. More recently, metadynamics combined with integrated tempering sampling was employed to estimate the free energy barrier and the nucleation rate of TIP4P/Ice at 230 K [40]. To achieve this, they used a new long-range collective variables S_X based on the specific features of X-ray diffraction pattern of ice and liquid to distinguish the two phases, coupled with pair entropy S_S [155]. The first one measure the ice formation, while the second one allow the system to explore the ice polymorphs. The resulting estimate of the nucleation rate was lower than the experimental one, albeit within its statistical error, and its global trend with respect to temperature was also in good agreement with experimental estimates. As for the previous studies on mW and TIP4P/Ice, the estimated critical cubicity is coherent even if with slightly higher fluctuation $C = 0.7 \pm 0.1$. However the critical nuclei size $N_c = 314 \pm 20$ is much lower than for the two previous studies ($N_c = 474 \pm 12$ in Ref. [42], and $N_c = 450 \pm 35$ in Ref. [38]), with a free energy barrier $\Delta G_c = 52 \pm 6\ k_B T$. This discrepancy may be due to the different choices of order parameter in the different works, as Ref. [38] showed that for the same configuration different order parameters yield a wide range of nucleus sizes.

The present work aims to produce reliable, high-quality transition path ensemble results, comparable to those in Ref. [38] based on the mW model, employing the more accurate TIP4P/Ice model. This will allow us to estimate rigorously critical cubicity and nucleus size, along with a quantitative analysis of the stacking disorder mechanism. It will also set the ground for more advanced methods of kinetics reconstruction based on Langevin or master equation models, the idea of these methods being to estimate the free energy landscape and the diffusion coefficient from a set of short shooting trajectories [43]. To this end, we used aimless shooting, both in the original version and with a modified shooting range approach, coupled with the PIV distance as order parameter to describe the nucleation reaction. As already discussed in the

chapter on the liquid-liquid transition, the PIV metric has been applied to a variety of systems, showing its strength and versatility. Here we will further assess its quality by comparing it with other order parameters using the rigorous maximum likelihood optimization scheme.

7.2 Generation of initial reactive trajectories

As discussed in section (5.2), to perform transition path sampling of ice nucleation we need an initial set of reactive trajectories connecting the liquid and the solid, and as said, several methods exist for this. We initially aimed to use metadynamics to generate such trajectories, extending the approach of Ref. [8], but after facing technical difficulties we switched to the simple and efficient seeding technique, exploiting the knowledge acquired with metadynamics simulations.

7.2.1 Freezing and melting with metadynamics

Guided by the results of the study made employing forward flux sampling in Ref. [42], we adopted a simulation box of $N = 4096$ water molecules described by the TIP4P/Ice inter-atomic potential. All the metadynamics simulations were performed under NPT condition at 230 K and 1 bar.

Order parameter

As for the liquid-liquid transition in the previous chapter, we used permutation invariant vectors (PIV) to define our order parameter able to distinguish liquid and ice. Here we coupled the PIV distance with the path collective variable, see (4.2.3) and (4.2.4) for their respective definitions. Here to define the PIV we only used Oxygen-Oxygen and Hydrogen-Hydrogen distances, which leads to two PIV blocks

$$V_{ij}^1 = w_1 \sigma \left(\left(\frac{V}{V_0} \right)^{1/3} r_{ij}^{OO} \right) \tag{7.2.1}$$

$$V_{ij}^2 = w_2 \sigma \left(\left(\frac{V}{V_0} \right)^{1/3} r_{ij}^{HH} \right) \tag{7.2.2}$$

where $r_{ij}^{\alpha\alpha}$ is the distance between atoms i and j of type $\alpha \in O, H$. The PIV blocks were weighted with $w_1 = 1$ and $w_2 = 0.2$. The switching function used is a rational one with the following formula

$$\sigma(r) = \frac{1 - (r/r_0)^4}{1 - (r/r_0)^{12}} \tag{7.2.3}$$

with $r_0 = 0.7$ nm, and is the same as in Ref. [8], where it was chosen to maximize the PIV distance between I_h and liquid.

To define the path collective variables, we used a liquid box L and a crystalline box of hexagonal ice I_h, both equilibrated at 230 K and 1 bar, so that S and Z are

$$S(X) = \frac{1 \times \mathrm{e}^{-\lambda D_{LX}} + 2 \times \mathrm{e}^{-\lambda D_{\mathrm{I_h} X}}}{\mathrm{e}^{-\lambda D_{LX}} + \mathrm{e}^{-\lambda D_{\mathrm{I_h} X}}} \tag{7.2.4}$$

$$Z(X) = -\frac{1}{\lambda} \log \left(\mathrm{e}^{-\lambda D_{LX}} + \mathrm{e}^{-\lambda D_{\mathrm{I_h} X}} \right) \tag{7.2.5}$$

where D_{LX} is the squared Euclidean PIV distance between liquid water and a configuration X, and $D_{\mathrm{I_h}X}$ is the same for hexagonal ice. We chose $\lambda = 0.05$ following the common rule of thumbs that $\lambda D_{L\mathrm{I_h}} \approx 2.3$.

Simulation settings

For all our simulations (metadynamics, seeding and transition path shooting) we employed the GROMACS 2018.3 simulation package [132]. We adopted a 2 fs timestep, short-range interactions were truncated at 0.90 nm and the particle mesh Ewald method was used to compute electrostatic interactions [69]. Bond constraints were maintained using the LINCS algorithm with a fourth order expansion [133].

To control the temperature we used the stochastic velocity rescaling thermostat with a relaxation time of 0.5 ps [65], while to control pressure we used an isotropic Parinello-Rahman barostat with a relaxation time of 2 ps [67].

In metadynamics simulations we deposited Gaussian hills of width $\sigma_S = 0.04$, $\sigma_Z = 0.4$ and height of 0.478 kcal/mol every ns in the space of path collective variables.

Attempts to freeze the liquid

The first issue we stepped on, was that if the calculation speed of the PIV at each timestep of the trajectory was acceptable for a small system of $N = 800$ molecules, it was not so for the larger system under study of $N = 4096$ molecules, as performance was divided by ~ 30. This is due to the $O(N^2)$ complexity of computing PIV and its derivatives in its plumed implementation. Even if base performance was multiplied by 4 after enhancement of the code, still on a typical super computer node it was hard to get more than 4 ns of simulation *by day*. It is important to remark than with the Q_6 collective variable the performance is almost one order of magnitude slower.

As a consequence, the poor computer performances limited the range of our simulations and the amount of trial-and-error cycles we could make. After simulations of more than 150 ns, no crystallization was in sight, despite a significant bias added to the liquid part of the free energy landscape. Probably this is due, at least in part, to the specific definition of the path collective variable S. Put it simply, our way of defining S based only on two reference states, the bulk liquid and the bulk crystal, roughly measures the proportion of crystalline structure in the system. Even if this was perfectly fine for the 800-molecules system of Ref. [8], as a small amount of crystalline structure translated in a significant proportion of the whole system, in a much larger system the collective variable cannot clearly resolve the early nucleation stage.

In fact, the PIV has difficulties in distinguishing a purely liquid state from a state with less than ~ 50 crystalline molecules arranged in a cluster for a box of $N = 4096$ molecules, as can be seen in figure 7.2.1. Unfortunately, this is specifically what would be required for the external bias to enhance the apparition of small clusters and for them to grow into larger one. It is possible that a definition of collective variables including intermediate ice cluster sizes, between the reference liquid and the reference bulk crystal, or replacing the bulk crystal with a supercritical nucleus as second reference structure, might allow to tackle the nucleation in large simulation boxes. This topic will be the subject of future investigations in our research group.

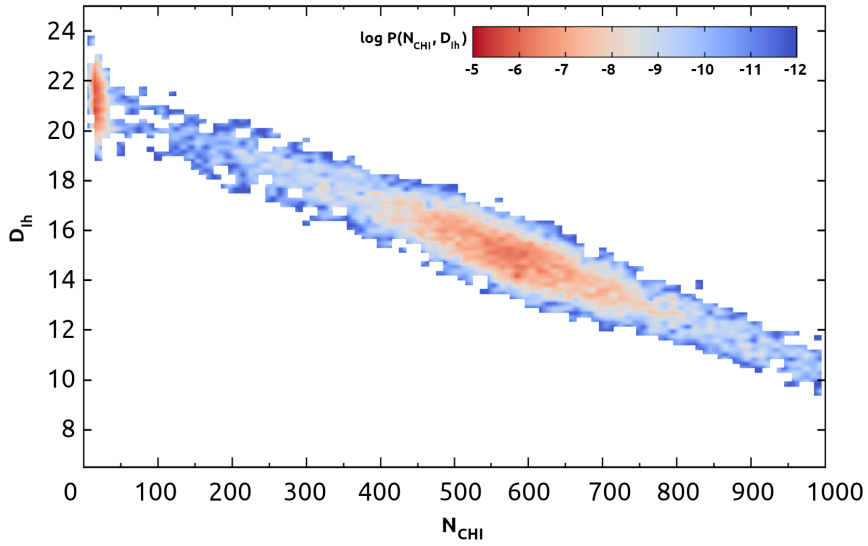
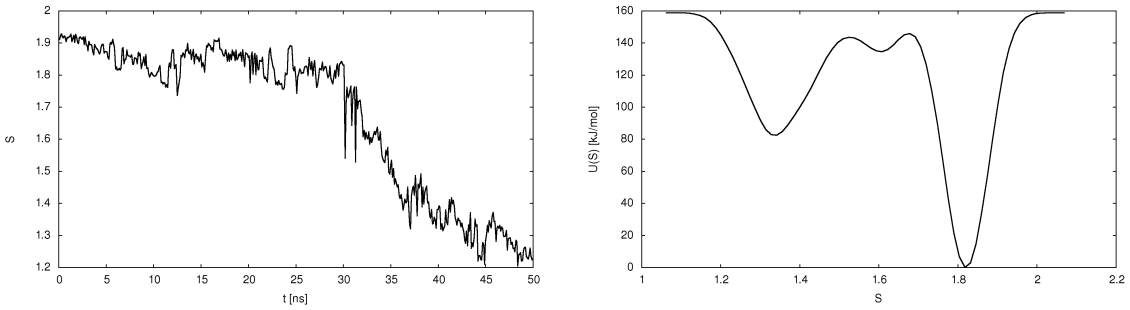


FIGURE 7.2.1 – Correlation between $D_{\mathrm{I_h}}$ and N_{CHI}, see sections (7.3.1) and (4.2.1) for their definition, from trajectories generated with hexagonal seed (see section (7.2.2)), colored according to $\log(P(N_{CHI}, D_{\mathrm{I_h}}))$. For $N_{CHI} < 50$ we lose the linear correlation between the two variables.

Melting a crystalline state



(a) (b)

FIGURE 7.2.2 – a) Time evolution of the PIV-based path collective variable S as defined in equation 7.2.4 during metadynamics simulation of melting. At $S \sim 1.9$ the system is a bulk crystal and at $S \sim 1.1$ a bulk liquid. b) Associated bias profile, an overestimation of the free energy profile, as a function of S for a fixed value of $Z = -2.1$, as in this case the Z coordinate did not resolve any specific features.

However, the current two-state definition of PIV-based path collective variables can properly describes the melting, as we were able to perform the transition in the reverse direction, starting from a bulk crystalline $\mathrm{I_h}$ state and ending in a liquid one. In this case, only 50 ns are required to melt our system, with the bias profile yielding a very rough (over)estimate of the free energy profile (since we did not observe reversible transitions). Figure 7.2.2 shows the generated reactive trajectory with the associated bias profile.

To summarize, it is effectively possible to generate some reactive trajectories by using metadynamics and PIV-based coordinates, albeit with a large wall-clock

time due to the burden of computing the PIV at each simulation timestep. As a consequence, few reactive trajectories can be generated in a limited time span. As it is always the case with enhanced sampling techniques based on bias potentials or forces, the transition states explored in this way need to be validated and possibly improved by means of committor analysis and path sampling techniques. As an effective alternative to metadynamics, in the following we generated reactive trajectories adopting the seeding technique, simple to implement and of small computational cost, since it lacks the need to compute the PIV at each timestep.

7.2.2 Exploration of the transition state with seeding

Choice of the optimal temperature

Following the path of our previous study with metadynamics, we first performed a seeding analysis at 230 K. On one hand, this temperature gives a critical nucleus size of $\sim 400 - 500$ as discussed in (7.1.4), which is small enough to prevent any perturbation due to the periodic boundary conditions in our box of 4096 molecules; on the other hand, the resulting dynamics is very slow. The transition path time, i.e., the typical time required in a reactive trajectory to connect liquid from ice, or the reverse, at this temperature is larger than 200 ns. As we will need to generate thousands of these trajectories to sample the transition path ensemble, it is of major importance to reduce the latter duration. This is why, based on previous seeding study that estimated the critical nuclei size for several temperature [36], we switched to 237 K. At this temperature the typical transition path time is of ~ 100 ns, while the critical nuclei of ~ 600 are still small enough to prevent periodic boundary effects due to the small size of our simulation box.

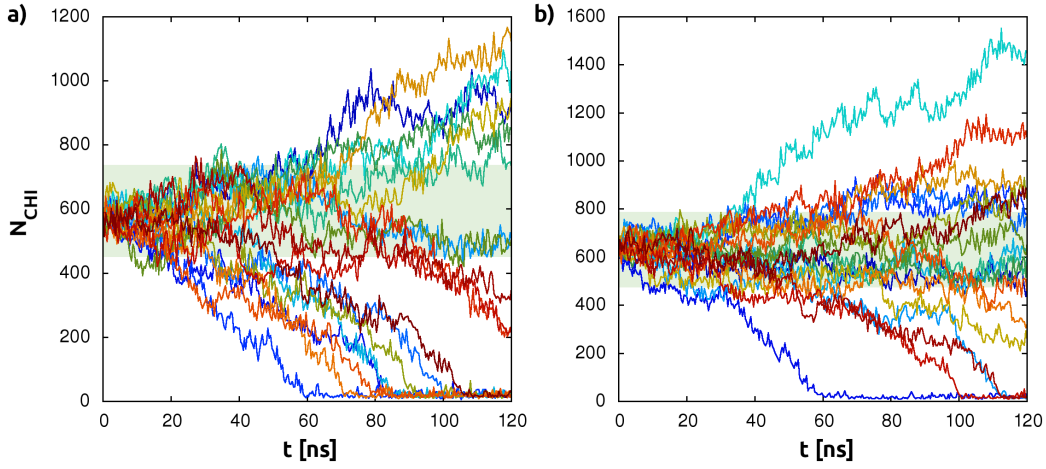


FIGURE 7.2.3 – Evolution of the largest nucleus size N_{CHI}, defined in section (4.2.1), for several shootings made from (a) a hexagonal and (b) a cubic nucleus seed at 237 K and 1 bar. Note the difference in vertical scale between the two. The light green area indicate roughly the transition region in which the system appears to diffuse rather freely on an almost flat free-energy landscape.

Seeding procedure

The main difficulty of seeding resides in the generation of the initial configurations. Cutting out a cluster of crystalline molecules from a bulk crystal and surrounding it with liquid molecules cut out from a bulk liquid is an easy task, but subsequent equilibration of the liquid-crystalline interface without spoiling the cluster seed can be non-trivial. Here we used the following procedure :

— extract a spherical nucleus of hexagonal or cubic ice from perfect crystalline ice boxes relaxed with the TIP4P/Ice potential at 237 K and 1 bar ;
— solvate the nucleus in a cubic box of liquid water (using gromacs tools) to get a total of 4096 molecules ;
— equilibrate the interface for 5 ns by restraining the positions of icy molecules only, using a strong harmonic bias with a spring constant of 400'000 kJ/mol to maintain RMSD distance close to zero with respect to the initial nucleus structure ;
— equilibrate the final seed state for 2 ns without any restraints.

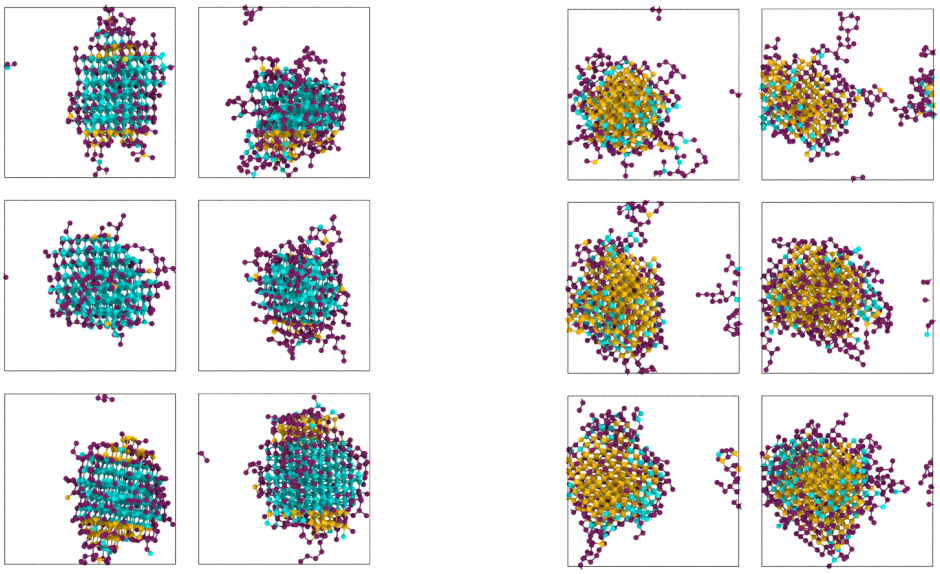
This last relaxation step let the nucleus adjust to the surrounding liquid, and it leads to slight changes in the exterior shell of the nucleus. For instance for I_c , during this equilibration a small I_h "shell" spontaneously forms around a part of the initial seed.

From each hexagonal or cubic seed configuration we shot a series of 40 unbiased trajectories, their initial momenta being drawn from the Boltzmann distribution at 237 K. By counting the ratio of shootings that end in a liquid state and of those that end in crystalline state, we can estimate approximately the committor value ϕ of the seed configuration. In our case we consider to have a transition state if $\phi = 0.5 \pm 0.2$, that is if we have a similar probability to evolve towards the liquid or crystalline state.

Spontaneous exploration of the transition region

By repeating the above procedure for several initial nuclei size, ranging from 400 to 900, we found the critical sizes $N_c \approx 590$ starting from I_h nucleus and that $N_c \approx 670$ starting from I_c nucleus. Besides allowing to estimate the critical size, this kind of trajectories provide important insight into the nucleation process and its detailed mechanism. Figure 7.2.3 shows that, remarkably, regardless of the initial nucleus structure being I_h or I_c , a part of the trajectories explore for a long duration the transition region, extending to the whole shooting duration of 120 ns in some cases. During this exploration, the nucleus spontaneously changes its structure, often evolving toward a slightly stacking disordered structure. Furthermore, during the growth of the nucleus, regardless of the initial structure, layers of cubic and hexagonal ice will spontaneously stack in a random way.

Figure 7.2.4 shows the end state of some shootings for which the largest nucleus size does not evolve much during the relaxation, but instead undergoes important structural change. Seeding from hexagonal ice results in the addition of layers of cubic ice, and reversely seeding from cubic ice results in added layers of hexagonal ice, in a disordered way. These observations provided the key to strongly enhance the efficiency of the transition path sampling protocol, as discussed in section (7.4).



(a) (b)

FIGURE 7.2.4 – Selected final nuclei after 120 ns of relaxation from (a) hexagonal or (b) cubic nucleus seeds, in the case where the final and initial nucleus sizes are similar ±100 molecules. Here we only represent the oxygen atoms in the largest cluster, omitting the liquid part and hydrogen for clarity. Using Chill+ to distinguish the type of the atoms (see section (4.1.3)), we use the following color code : blue represent hexagonal ice, yellow cubic ice and purple interfacial ice.

7.3 Order parameter quality

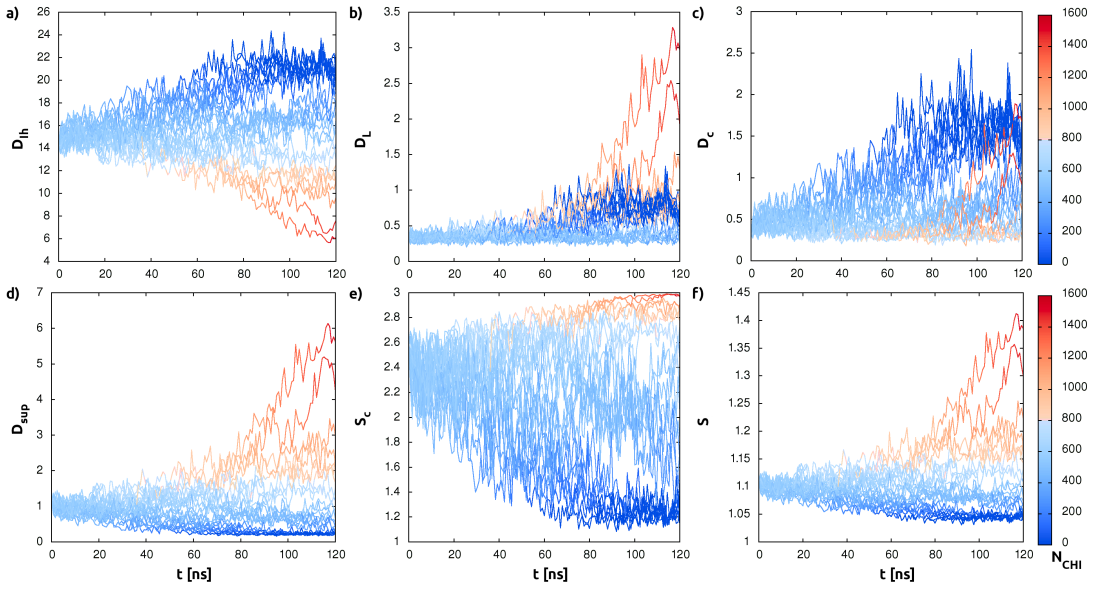


FIGURE 7.3.1 – Evolution of several collective variables for shootings made from a hexagonal nucleus seed. The color code indicate the corresponding largest nuclei size, dark blue being $N_{CHI} = 0$ and red $N_{CHI} = 1800$. a) D_{Ih}, b) D_L, c) D_c, d) D_{sup}, e) S_c, f) S.

7.3.1 Choice of the order parameter

The information contained in seeding trajectories allowed us to compare different possible definitions of PIV-based collective variables to choose the most promising one for further work. The first thing we observed was that considering only the oxygen atoms, i.e., neglecting the information about hydrogen positions, only changed the PIV distance by a constant factor and substracted no important information,so all of the subsequent analysis are made with PIV that consist of one Oxygen-Oxygen block

$$\mathbf{V}_X = \text{sort}\left\{\sigma\left(\sqrt[3]{\frac{V}{V_0}}r_{ij}^{OO}\right)\right\} \tag{7.3.1}$$

with the same switching function as defined in equation 7.2.3. Based on this definition of the PIV, we considered several possible collective variables :

$$\begin{aligned}
D_L &= |\mathbf{V}_X - \mathbf{V}_L|^2 \\
D_{Ih} &= |\mathbf{V}_X - \mathbf{V}_{\mathrm{I_h}}|^2 \\
D_c &= |\mathbf{V}_X - \mathbf{V}_c|^2 \qquad D_{sub} = |\mathbf{V}_X - \mathbf{V}_{sub}|^2 \qquad D_{sup} = |\mathbf{V}_X - \mathbf{V}_{sup}|^2 \\
S &= (\mathrm{e}^{-\lambda D_{Ih}} + 2\mathrm{e}^{-\lambda D_L})/(\mathrm{e}^{-\lambda D_{Ih}} + \mathrm{e}^{-\lambda D_L}) \\
S_c &= (\mathrm{e}^{-\lambda D_{sub}} + 2\mathrm{e}^{-\lambda D_c} + 3\mathrm{e}^{-\lambda D_{sup}})/(\mathrm{e}^{-\lambda D_{sub}} + \mathrm{e}^{-\lambda D_c} + \mathrm{e}^{-\lambda D_{sup}})
\end{aligned}$$

where the $\mathrm{I_h}$ subscript means that we use a hexagonal reference structure, c a critical one and L a liquid one. *sub* and *sup* indicate specifically chosen states which are equidistant to the critical reference structure and verify that $|\mathbf{V}_c - \mathbf{V}_{sub}|^2 \simeq |\mathbf{V}_c - \mathbf{V}_{sup}|^2$. Physically they correspond to sub-critical and super-critical state, that is a box where the largest nuclei size is below or above the critical one.

Figure 7.3.1 show the trajectories projected on all of the previous collective variables. From inspection of the graphs it appears that only three definitions offer a proper separation of the two phases, D_{sup}, S, S_c and D_{Ih}. Among them D_{Ih} seems to yield the best result, as it distinguishes neatly the transition state and offers a somewhat symmetric space with respect to the liquid and to ice. This rather qualitative observation will be quantitatively assessed later on when we will use the maximum likelihood optimization scheme in section (7.3.3).

As already discussed in section (5.2), it's important to note that the two transition path sampling methods that we will use have not the same requirement on the collective variable. If for aimless shooting we only need a collective variable able to distinguish the two end states, here liquid and crystalline, for aimless shooting within a range we also need the variable to be able to identify quite precisely the transition state region, as we will need to define a shooting range in collective variable space matching such region in order to obtain an efficient sampling.

As a first indication, figure 7.3.2 shows the almost-linear correlation between D_{Ih} and the more traditional and widespread definition of largest nucleus size N_{CHI}, which is the sum of the number of cubic, hexagonal and interfacial ice molecules in the largest cluster, where the molecules are identified using Chill+ algorithm [9, 38], see section (4.1.3) and (4.2.1). Those are computed from the path ensemble sampled by shooting range aimless shooting. Please note that here we present this extensive set of configurations in the interest of having optimal statistics, but that the same analysis was performed beforehand with the more restricted dataset of

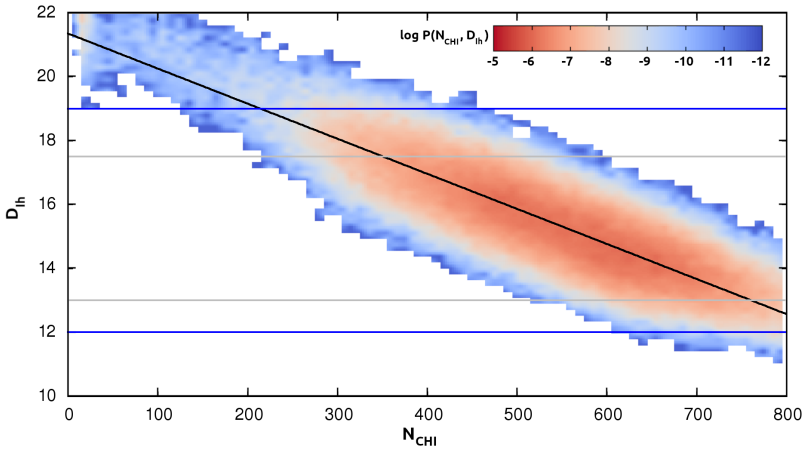


FIGURE 7.3.2 – Correlation between D_{Ih} and N_{CHI} from accepted trajectories generated with the shooting range algorithm seeded with hexagonal ice, colored according to $\log(P(N_{CHI}, D_{Ih}))$. The black line is a simple linear regression. Blue lines indicate D_{Ih} thresholds to assign the future evolution of the system to liquid or ice, and gray lines delimit the region with highest acceptance rate for the shooting range algorithm.

seeding trajectories (the same remark holds for the following paragraph). Regarding the correlation between the two collective variables, an important point to be aware of is that a given value for the largest cluster size can lead to a large variety of nucleus structures : some may be very compact and spherical, some more elongated, some may have more interfacial ice than cubic or hexagonal ice, and some may be composed of almost disconnected blocks. This variability explains the breadth of the distribution in figure 7.3.2, notwithstanding the good degree of linear correlation.

To identify trajectories committed to ice or liquid water during the execution of the shooting range algorithm, we need to define two threshold values of the D_{Ih} variable : the transition state region will span the enclosed interval. Figure 7.3.3 shows the distribution of the size of the largest nucleus at different fixed values of D_{Ih}. These distributions have approximately Gaussian shape, with negligible or no overlap even for the smallest separation between the two thresholds. As we can see, the center of the distribution moves linearly toward higher crystallite size when we lower the D_{Ih} value, which is what we expect due to the linear correlation between the two variables. Based on this plot and on the previous seeding data, we tentatively chose the threshold values of 12 for ice state and 19 for liquid state, those are indicated by blue line on figure 7.3.2. A compelling assessment of such thresholds clearly requires committor analysis, as discussed in the next section.

7.3.2 committor analysis

To further test the quality of the thresholds used to distinguish trajectories committed to liquid and ice, we performed a simple committor analysis. From reactive trajectories generated with the shooting range algotihm starting from a hexagonal seed, we selected a set of 4 configurations at the liquid boundary, with $D_{Ih} = 19$ and largest nucleus size of 300, 350, 400, 450. We evolved $n = 15$ independent trajectories for 200 ns from each of these states, drawing their initial momenta from the Boltzmann distribution. Almost all of them evolved toward their predicted final phase,

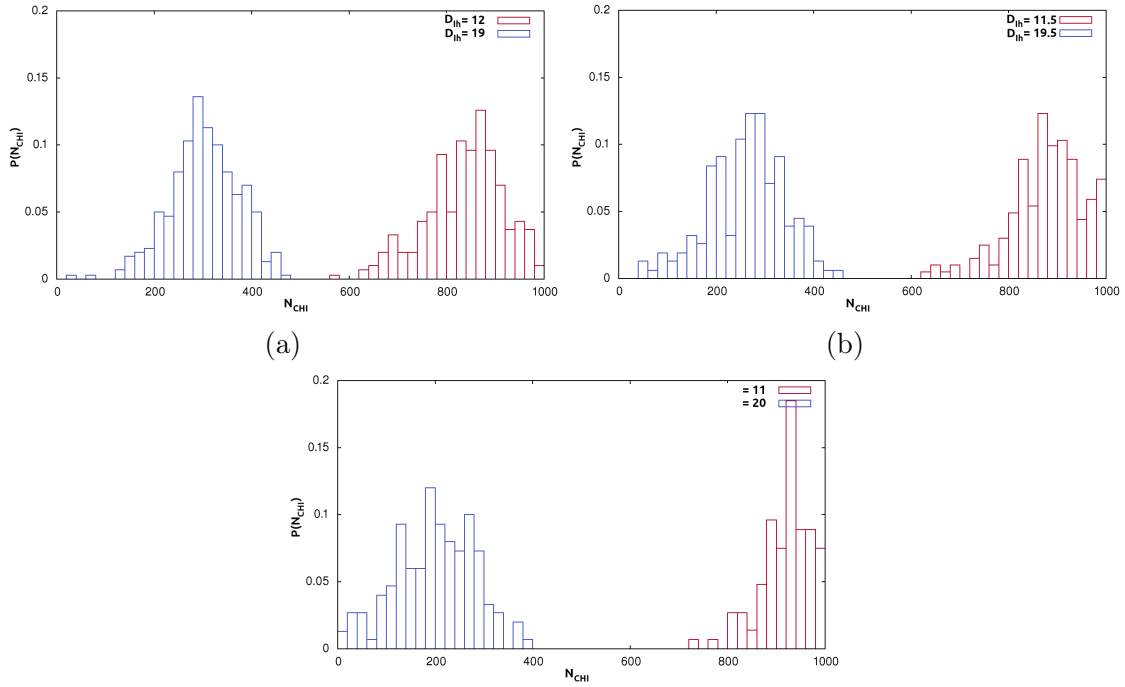


FIGURE 7.3.3 – Probability distribution of the size of the largest nucleus N_{CHI} at given fixed values of the PIV distance from hexagonal ice D_{Ih}. a) $D_{Ih} = 12$ or 19, b) $D_{Ih} = 11.5$ or 19.5, c) $D_{Ih} = 11$ or 20, using reactive trajectories sampled with aimless shooting within a range starting from a hexagonal seed.

liquid water. Only 3 trajectories shot from $N_{CHI} = 450$ evolved toward a crystalline state.

In the same way, we selected a set of 4 configurations at the ice boundary, with $D_{Ih} = 12$ and largest nucleus size of 650, 700, 750, 800. Shooting $n = 15$ trajectories from each of these states in the same way as for the liquid boundary, we found again that almost all of them evolved toward a crystalline state, as expected. Only 6 trajectories shot from $N_{CHI} = 450$ evolved toward a liquid state. By weighting the probability of the initial configurations using the distributions in figure 7.3.3, we find less than 1% of falsely attributed end-states for both boundaries. This analysis confirms the appropriateness of D_{Ih} as collective variable definition, and it demonstrates that $D_{Ih} = 12$ and 19 are valid thresholds to predict whether a trajectory will irreversibly evolve towards ice and liquid water, respectively.

7.3.3 Maximum likelihood optimization

Finally, we quantitatively assessed the quality of our PIV distance-based collective variables by using the maximum likelihood optimization scheme [88] described in section (4.3.1). For this we used the transition state ensemble generated with the mW water potential in Ref. [38], kindly provided by the Authors. The ensemble was generated with very extensive aimless shooting simulations at 230 K and 1 bar and it includes 21523 atomic configurations of 4608 molecules. We compared the many collective variables analyzed in Ref. [38] with all the ones we introduced in section

(7.3.1). We are not going to present in detail all of the variables of Ref. [38], but globally they can be grouped into three classes :

— Size variables named $N_{something}$, that measure the crystallite size, where *something* is some criterion used to distinguish icy molecules from liquide one.

— Energy variable named $E_{something}$, which is the sum of all potential energy of water molecules in the cluster, again using *something* to distinguish liquid and ice.

— Structural variables, that describe internal structure or shape of the crystallite. Namely, there are the cubicity C and the gyration radius R_g.

Remarkably, the collective variable that we chose to perform our path sampling simulations based on the analyses in the previous sections, D_{Ih}, achieves the highest score in the long list of candidate variables, as shown in table 7.3.1.

RC	$-\ln L$	$\Delta\ln L/BIC$	RC	$-\ln L$	$\Delta\ln L/BIC$
D_{Ih}	10466	0	E_I	14222	-724
$N_{CHI-\langle Q_6\rangle}$	11314	-163	N_I	14233	-726
N_{CHIn_4}	11342	-168	$N_{E-pp<11.5}$	14010	-683
$N_{CHI-\langle Q_6\rangle+solv}$	11369	-174	$N_{Q_6>0.57}$	14501	-777
$N_{CHI+solv}$	11376	-175	$N_{E-pp<11.6}$	15841	-1035
E_{CH}	11426	-185	n_4	16411	-1145
N_{CH}	11434	-187	$N_{Q_6>0.5}$	16535	-1189
E_{CHI}	11554	-210	N_H	17712	-1396
N_{CHI}	11573	-213	E_H	17714	-1397
$N_{Q_3>0.7}$	12279	-349	E_C	18510	-1550
Rg^2	12401	-373	N_C	18512	-1550
$N_{Q_6>0.55}$	12715	-433	C	19720	-1783
$N_{E-pp<11.3}$	12796	-449			
$N_{E-pp<11.4}$	13141	-515			
N_{solv}	13331	-552			

TABLE 7.3.1 – Maximum likelihood value and score for all collective variables defined in Ref. [38], with the addition of D_{Ih}, which scores the best. The dataset includes 21523 transition state structures explored with aimless shooting simulations using the mW potential.

It is interesting to note that the most widespread criterion for assessing the quality of a reaction coordinate is based on information from committor analysis, as in the technique here above. Indeed, the committor function is commonly considered "the" ideal reaction coordinate for any transition process between metastable states, from crystal nucleation to protein folding to chemical reactions [11, 86, 89, 88]. However, in high-barrier transition processes like ice nucleation such information is available only within a limited energy range (a few k_BT units) close to the barrier top, due to the numerical difficulties involved in sampling committor values very close to zero or to one. This implies that there is ample room to develop approaches alternative to direct committor estimation to investigate the quality of reaction coordinate at the early stage of nucleation (close to the liquid state) and the early stage of crystal

fusion. Future efforts in this direction are desirable, and could take advantage of the availability of a large data set of unbiased transition pathways like those sampled in this work.

7.4 Sampling of the transition path ensemble

Now that we have defined our collective variables the liquid and crystalline state basins associated, we can start to sample the transition path ensemble. We stress here that, due to the need of generating thousands of MD trajectories of hundreds of nanoseconds of duration, such endeavor implies a massive amount of computer resources, estimated in our case to be of more than 6 million CPU hours (partly obtained through a PRACE European grant), which probably explains why no such attempt has been previously reported in the literature.

In a first attempt, following the study in Ref. [38] with the mW model of water, we adopted the standard aimless shooting approach [12, 13]. However, due to the low sampling efficiency of this method and based on the specific features of nucleation trajectories in the transition state region, we switched to a variant of the algorithm, i.e., aimless shooting within a range [41], that resulted significantly faster and proved a key tool to tackle our ambitious goal.

7.4.1 Standard aimless shooting

So the first transition path sampling procedure that we use is aimless shooting, that we presented in section (5.2.2). For this algorithm we only need to define two end-state basins (regions in phase space corresponding to local free-energy minima), here liquid and ice, based on an order parameter, here D_{Ih} as discussed. We started several independent aimless shooting simulations from a hexagonal seed containing $N_{CHI} = 590$ water molecules. As an initial reactive trajectory, we concatenated two seeding trajectories at 237 K and 1 bar : one evolving towards the liquid (seed melting) and the other evolving towards the crystal (seed growing).

Based on figure 7.2.3.a, we estimated that a maximal duration of 60 ns for the backward and forward shootings would be enough to reach the liquid or crystal thresholds in order parameter space – hence the corresponding basins – for the majority of the trajectories. This choice is obviously linked to performance issues, as in a world with infinite resources we would have avoided to introduce a maximal duration in order to avoid "undecided" trajectories, not connecting two basins (possibly the same) during the relaxation.

To achieve a better statistical description of nucleation, which is an intrinsically stochastic process depending on the random walk (diffusion) of thousands of water molecules, we launched 25 independent aimless shooting simulations, with a time step for sampling potential transition states of $\Delta t = 0.2ns$, see section (5.2.2). In this scheme each run was able to achieve ~ 1.2 steps per day, depending on the super computer availability. It is important to stress that, contrary to techniques like umbrella sampling, where tens of independent trajectories (the different windows) can be executed concurrently (at the same time), each one of them exploiting parallelism, aimless shooting is intrinsically sequential, since the result of each pair of forward and backward trajectory (i.e., to which basin they are committed) needs to be known to execute the following trajectories, each single trajectory of course exploiting pa-

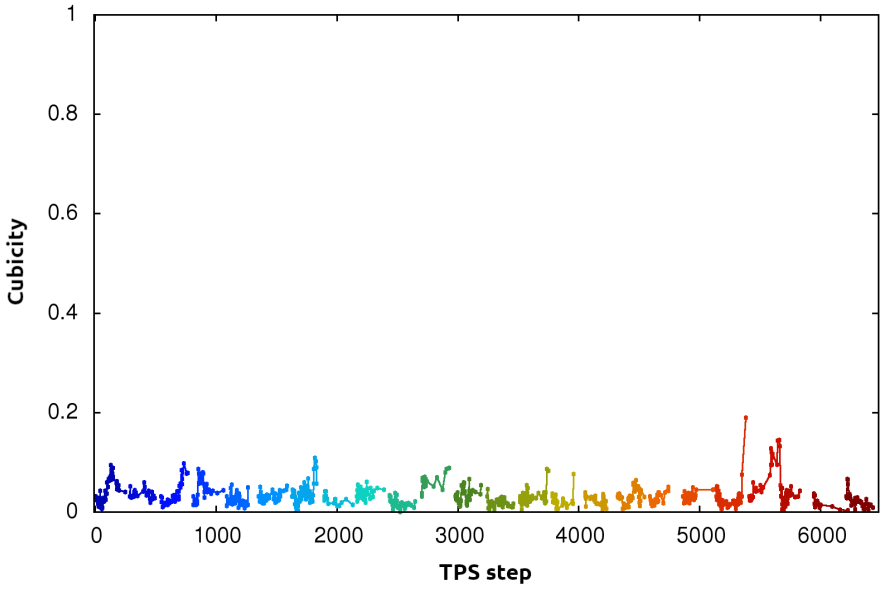


FIGURE 7.4.1 – Evolution of the cubicity for aimless shooting simulations starting from hexagonal ice. Each color represents an independent simulation, concatenated one after another for easier visualization.

rallelism. The crucial difference between the two types of simulations is therefore the much longer wall-clock duration of aimless shooting, for a same cumulative amount of computer resources employed. In the specific case of PIV-based collective variables, another difference between the two classes of simulations is the need to compute the expensive variable at each timestep in umbrella sampling (due to need to apply biasing forces along the trajectory), whereas the variable can be computed infrequently in aimless shooting (for example, every several thousands steps). In practice, even though we ran concurrently 25 aimless shooting simulations to improve statistics, sampling more than 6000 trajectories (out of which $\sim$ 12% connect liquid and ice) took approximately 7 months.

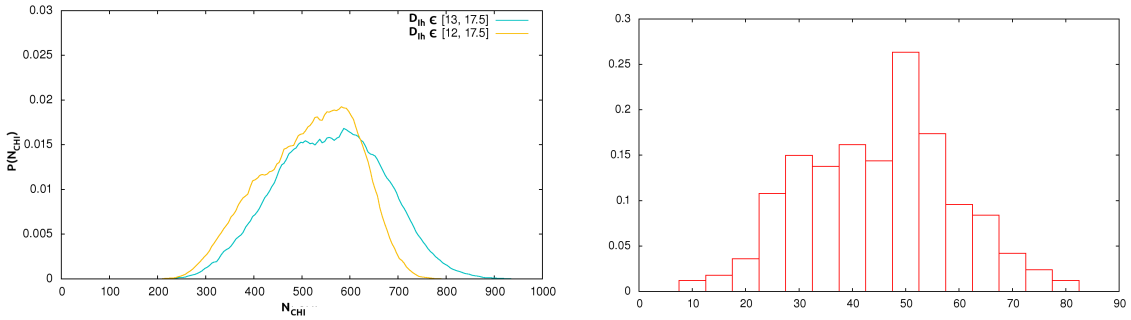
Figure 7.4.1 shows that the initial hexagonal seed evolve slowly, with some simulations displaying transition states transforming toward a slightly stacking disordered crystallite. Even if these data seem to follow the expected trend of increasing cubicity [42, 38, 40], it is hard to draw any conclusion from them, due to the relatively small amount of sampled points, having 38 accepted transition state per simulation in average, for more than 875 in total.

These results, coupled to the analysis of seeding trajectories (section 7.2.2), led us to adopt another transition path sampling technique. As shown in figures 7.2.4a and 7.2.3.a, during the seeding procedure we observed that the system could spontaneously explore a broad portion of phase space in the transition states region, since for a rather stable largest nucleus size N_{CHI} the cubicity was often observed to drift from ~ 0 to ~ 0.2 within a single trajectory of the order of 100 ns. In stark contrast with this spontaneous behavior, aimless shooting simulations require more than 200 Monte Carlo moves to observe a comparable evolution in the critical nucleus structure, due to the sub-nanosecond timestep separating potential transition state structures randomly sampled from the previous trajectory at each move. As discussed in (5.2.2), this is due to the design of the algorithm, which is best fitted to sample a narrow

phase space region.

7.4.2 Aimless shooting within a range

So we switched to the second transition path procedure called "aimless shooting within a range" [41], that we presented in section 5.2.3. The algorithm is basically the same as standard aimless shooting, except that we are not limited to a constant time step to randomly pick the next shooting point before or after the last accepted one. Instead, new shooting points are randomly sampled from a pre-defined shooting range in our collective variable space, allowing to explore very rapidly the top of the free-energy barrier – very broad in the case of ice nucleation.



(a) (b)

FIGURE 7.4.2 – (a) Distribution of the largest crystallite size in the sampling area of the shooting range algorithm, starting from hexagonal or cubic ice seed (blue or yellow respectively) from accepted trajectories. The distribution was estimated from the combined data set of all trajectories sampled with the shooting range algorithm. (b) Distribution of the half duration of accepted trajectories, estimated with combined data set of all trajectories obtained with shooting range algorithm starting from hexagonal ice.

We performed two sets of simulations, both at 237 K and 1 bar, differing by their initialization. The first set started from reactive trajectories connecting liquid and hexagonal ice, taken randomly from the previous standard aimless shooting simulations described in section (7.4.1), so that each one is initialized with a different reactive trajectory. The second set started from a reactive trajectory obtained from a cubic seed containing $N_{CHI} = 675$ water molecules, constructed and equilibrated with the same protocol adopted for the hexagonal seed.

For hexagonal ice we adopted the threshold values described in section (7.3.1), i.e., a trajectory is considered committed to the liquid when $D_{Ih} > 19$ and to the ice when $D_{Ih} < 12$. For cubic ice, values are chosen with the same procedure but they result slightly shifted to account for the lower stability of the crystallite ($D_{Ih} > 19$ and $D_{Ih} < 11$ for liquid and ice respectively). To properly choose the shooting range (which can be smaller than the region delimited by the committor thresholds) for hexagonal ice, we started by taking three different ranges with 5 independent runs each, to make a quick estimate of the acceptance ratio :

— $D_{Ih} \in [13, 17.5]$ (9.9 % acceptance)
— $D_{Ih} \in [13, 5 : 18]$ (7.9 % acceptance)
— $D_{Ih} \in [13, 18.5]$ (8.2 % acceptance)

and we kept the one with the highest acceptance ratio. For cubic ice we slightly decreased the lower bound to take into account the difference in stability, so that $D_{Ih} \in [12, 17.5]$. Figure 7.4.2a show the distribution of N_{CHI} in these shooting ranges.

We ran 15 independent sets of simulations with a time step $\Delta t = 0.2$ ns for hexagonal or cubic initial seed. Contrary to standard aimless shooting, where acceptance rate is directly linked to the time step, here it has less importance and should just be not too small nor too big. Initially we kept the same 60 ns duration as for the standard aimless shooting, but we rapidly switched to 80 ns to reduce the amount of "undecided" trajectories, that did not connect two stable states (liquid or ice).

Figure 7.4.2b shows the distribution of the accepted trajectories length, which allow us to properly quantify the effect of this cut-off choice. As you may see, it cuts the tail of our distribution, which amount for less than 1% of the total set. So far, we sampled more than 3800 trajectories that start from the hexagonal seed, of which $\sim$ 10% connect liquid and ice (transition paths). Starting from the cubic seed we sampled more than 1900 trajectories, of which $\sim$ 11% connect liquid and ice.

7.4.3 Difference of efficiency between the two techniques

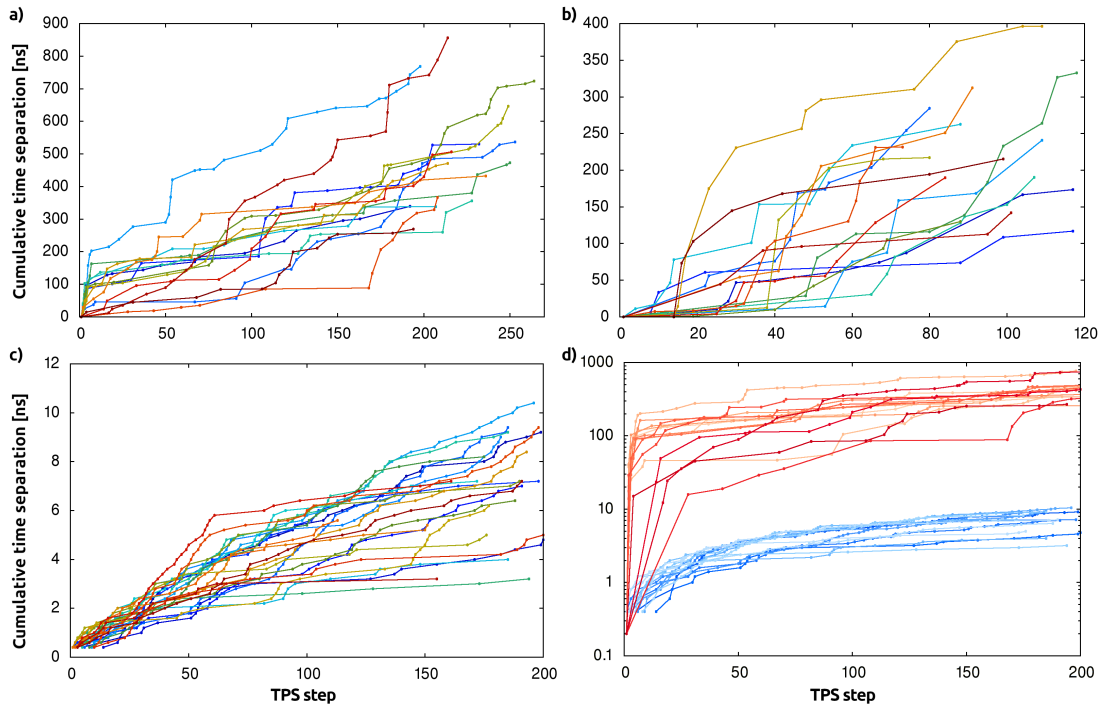


FIGURE 7.4.3 – Evolution of the cumulative time separation between the first and last accepted shooting points in transition path sampling simulations for a) aimless shooting within a range starting from I_h seed, b) aimless shooting within a range starting from I_c seed, c) standard aimless shooting starting from I_h seed. Note the difference in vertical scale between a), b) and c). Panel d) shows a comparison of the two methods, with standard aimless shooting in blue and aimless shooting within a range in red.

The reason why we passed from one transition path sampling technique to another is simple : aimless shooting within a range is a much more efficient algorithm

for the process under consideration, that we picture as slow diffusion on a smooth and broad free energy barrier starting from a sub-optimal region of the separatrix (such a picture holds in Ref. [38] from a direct estimation of the saddle region in the two-dimensional free-energy landscape as a function of nucleus size and cubicity with the mW model at 230 K). Figure 7.4.3 shows the cumulative temporal distance between the first and last accepted transition states of standard aimless shooting or its variant within a range, which provides a clear indication of the decorrelation of transition state structures during the sampling. It is clear that standard aimless shooting requires more than 100 steps to achieve the same time separation – hence decorrelation – as for its variant within a range in *one single* step on average. A similar conclusion can be drawn by comparing the speed of evolution of the cubicity of critical nuclei in Figures 7.4.1 and 7.4.4 (see the discussion in the next section).

Note that this is not expected to be true for all activated processes : in the present situation this efficiency is linked to the easy, spontaneous exploration of large regions of configuration space in the transition state domain, as observed during seeding. Probably, this behavior is a result of small diffusion coefficient combined with sizable average force pointing from the bad transition states towards the best ones. In practice, the aimless shooting within a range algorithm is able to pick successive shooting points separated by a very large time distance, often already decorrelated. While standard aimless shooting is forced to choose closely related shooting points, separated by a tiny temporal and structural distance.

7.4.4 Critical nuclei evolution

Figure 7.4.4 shows that the hexagonal seed has a strong tendency to evolve toward a stacking disordered crystallite, while for the cubic seed – notwithstanding the smaller number of path sampling steps so far – no strong drift is evident, probably due to the proximity between the initial cubicity and the expected optimal one (based on previous works [42, 38, 40]).

It is too early to draw conclusions about the precise value of the optimal critical cubicity (i.e., the one lying on the minimum free-energy path for nucleation), however the behavior of the two sets of simulations is consistent with $C \in [0.55, 0.8]$. At the time of writing, simulations are being completed in order to obtain a final estimate. Comparison of Figure 7.4.4 with Figure 1 in supporting information of Ref. [38] for (standard) aimless shooting simulations with the mW potential clearly indicates a similar behavior, suggesting that our simulations should soon fluctuate around an optimal critical cubicity.

Figure 7.4.5 shows the structural evolution of the critical nucleus within one run of aimless shooting within a range, starting from either a cubic or hexagonal seed, illustrating the evolution toward stacking disordered structures. As discussed in section (7.1.3), due to differences in symmetry, only the two basal planes of hexagonal ice can form coherent bonds with cubic ice, whereas all the four (111) planes of cubic ice can form coherent bonds with hexagonal ice. This explains why cubic ice tends to form one or two layers on the top and bottom of the hexagonal nucleus, whereas hexagonal ice tends to form "shells" around the initial cubic seed, with isolated hexagonal-like molecules disseminated on the (111) planes.

Finally, Figure 7.4.5 displays the set of all the critical nuclei obtained at the end of aimless shooting within a range runs, starting from either a cubic or hexagonal

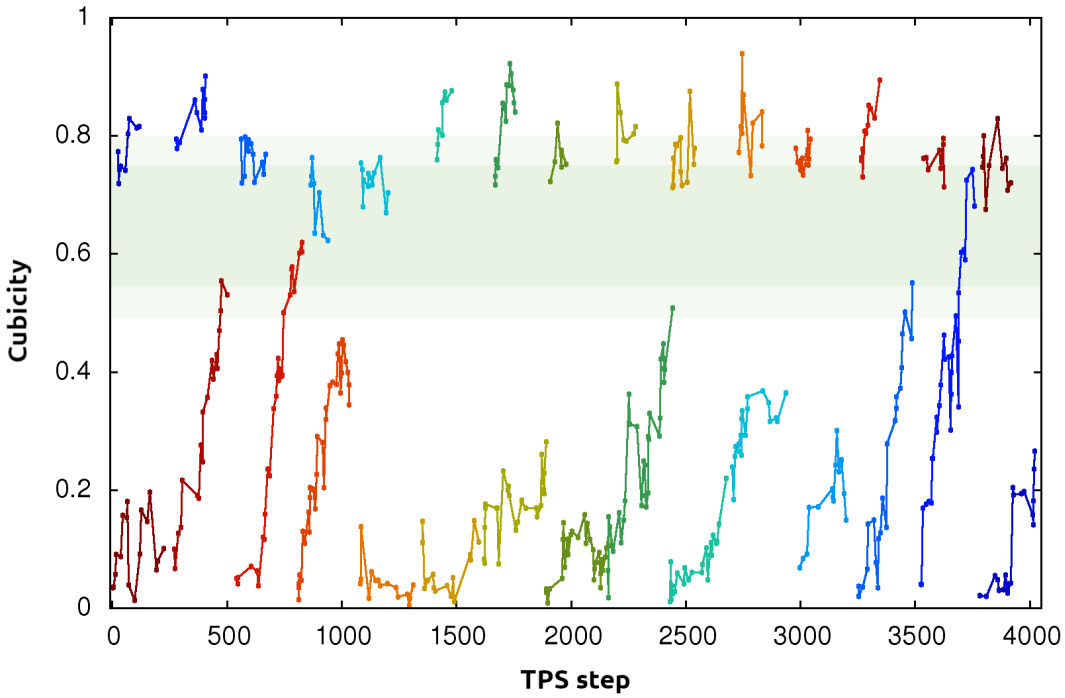


FIGURE 7.4.4 – Evolution of the cubicity for all shooting range simulations, starting from hexagonal ice or cubic seeds. Each colour represent an independent run, the different runs being shown one after another for ease of visualization.

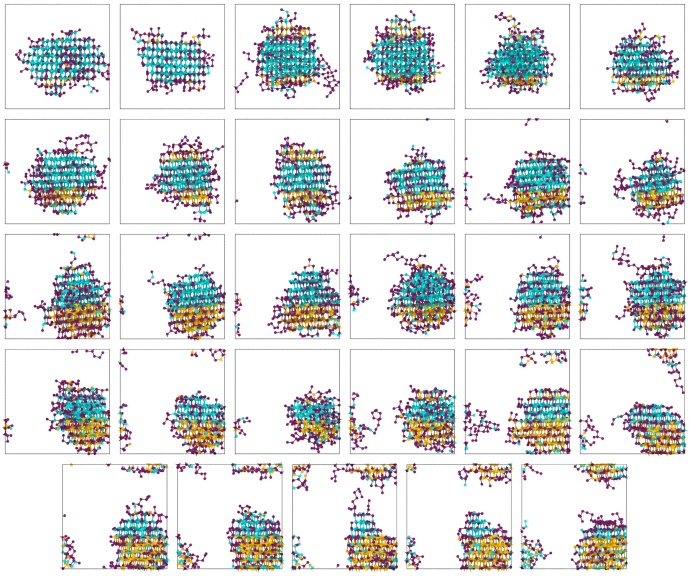
seed. In all cases, evolution towards stacking disorder features the same properties as discussed above. Even if simulations still need to be completed to determine the precise features of optimal critical nuclei, averaging over all the transition states we obtained, we find an average critical size of $N_{CHI} = 507 \pm 165$. This wide range of values is coherent with the study of Ref. [38] based on mW water. This is also coherent with the value of $N_c = 588$ found in Ref. [36] with seeding at 238.7 K.

Stacking mechanism

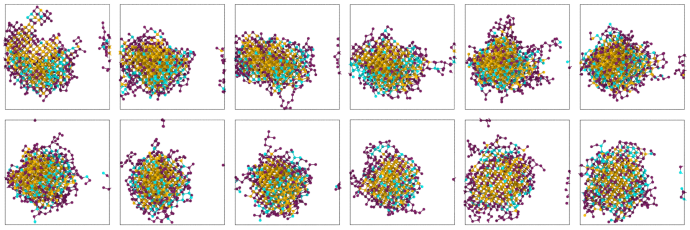
One may wonder how a nucleus can pass from being purely composed of hexagonal or cubic ice to a stacking disordered one with a high cubicity, despite I_h being the real stable phase.

Here we have performed a simple analysis to understand the mechanism that leads to the formation of disordered stacks of layers. For all of our transition paths sampled, every 50 ps we computed the state (liquid, cubic, hexagonal or interfacial) of oxygen atoms using Chill+. Then we counted the transition between atom states, e.g., how many times an oxygen atom goes from the cubic ice state to the liquid state, for instance. this procedure allows to track down the sources (previous states) of a specific state, and their relative fraction.

Figure 7.4.7 shows the evolution of the sources of cubic or hexagonal ice included into the largest crystalline cluster. Without ambiguity, the analysis shows that crystal-to-crystal structural change within the nucleus is marginal, as there is almost no direct transition between hexagonal and cubic states. This also shows that addition of new I_h molecule to the largest nucleus is mostly a two-step process, as molecules first



(a)



(b)

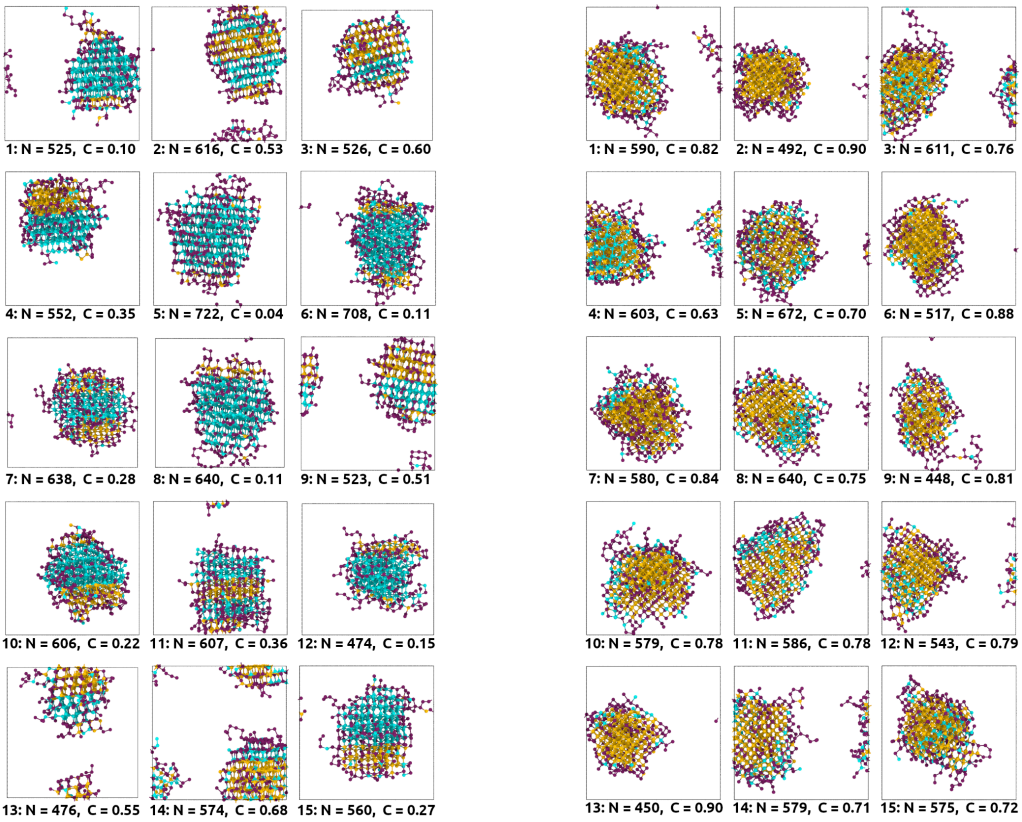
FIGURE 7.4.5 – Evolution of transition states sampled with aimless shooting within a range for two independent runs, starting from (a) an hexagonal seed and (b) a cubic seed, respectively. The representation is the same as in figure 7.2.4, i.e., we only represent oxygen atoms in the largest cluster, identifying their structure with Chill+ algorithm. Blue atoms are hexagonal ice, yellow cubic ice and purple are interfacial ice.

need to rearrange themselves into interfacial ice before assuming their final hexagonal configuration. On the contrary, addition of new I_c molecule is half of the times a one-step process, as molecules can directly go from a liquid to a cubic state.

It is interesting to note that this behavior remains valid regardless of the evolution of the cubicity, i.e., regardless of the progress of the transition path sampling simulation. It is coherent with the higher symmetry of cubic ice, see section (7.1.3), which lead to easier formation of coherent bonds with crystalline molecule that are already arranged in the largest nucleus. And it could be related to the higher intrinsic probability of liquid water molecules to adopt hydrogen bond patterns forming dihedral angles that are similar to those of cubic ice, as recently suggested from the analysis of TIP4P/ice simulations in Ref. [156].

7.5 Conclusions and outlook

The present study addressed an ambitious task : reconstructing for the first time the transition path ensemble for the accurate TIP4P/Ice model of water employing



(a) (b)

FIGURE 7.4.6 – Last shooting point sampled with aimless shooting within a range for all the independent run, starting from (a) hexagonal seed or (b) cubic seed. The representation is the same as in figure 7.2.4 and 7.4.5, i.e. we only represent oxygen atoms in the largest cluster, identifying their structure with the Chill+ algorithm. Blue atoms belong to hexagonal ice, yellow to cubic ice and purple to interfacial ice. The nucleus size and cubicity are indicated below each structure, whose ordering is coherent with Fig. 7.4.4.

a rigorous path sampling technique, i.e., aimless shooting. The latter technique was previously applied only to a coarse-grained model of water, mW, due to the very elevated computational cost connected to the need to generate thousands of transition pathways. In fact, two factors render these simulations very demanding : the transition path time often exceeds 100 ns at 237 K, while the need to accommodate the critical nucleus (including up to about 700 water molecules at this temperature) demands a simulation box containing thousands of molecules.

Two tools had a crucial importance in reaching our goal. The first is an improved aimless shooting algorithm that was developed recently [41] and that we adopted based on insight from seeding simulations, yielding very efficient exploration of the transition state ensemble for disordered ice nuclei. The second is the PIV-based topological metric [7, 84, 8], that allowed to precisely track the structural evolution of ice nuclei displaying a complex range of different structures, thanks to an excellent correlation with the committor function, i.e., the ideal reaction coordinate (see below).

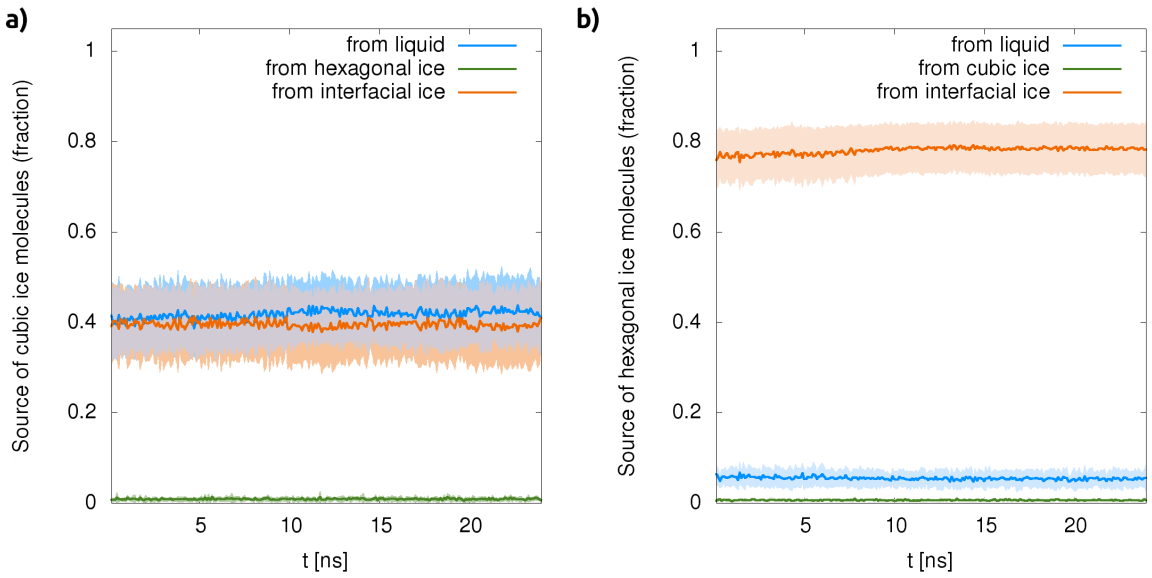


FIGURE 7.4.7 – Instantaneous fraction of new ice-like molecules of (a) cubic or (b) hexagonal type that evolved in the last 50 ps from the liquid, interfacial ice, or the opposite ice polymorph. The calculation is performed on a time span of 25 ns over the largest nucleus, averaging over the whole set of accepted transition paths collected from all the independent path sampling runs started from a hexagonal seed. The colored areas represent standard deviations.

Compared to previous studies, in our work

— we did not resort to a coarse-grained model like mW [38], reproducing many structural and energetic experimental properties but suffering from an incorrect description of kinetic properties ;

— we did not apply any external biasing forces [40], strongly altering the time scale of the process and leading to transition states that depend on the quality of the collective variables on which they act ;

— we employed a robust aimless shooting technique where full transition pathways are generated in a statistically correct way, avoiding the drawbacks of the forward flux sampling scheme where short segments of pathways are generated in a discretized collective variable space [42], leading to possible severe error propagation in the estimation of rate constants [37] ;

Our results provide a detailed picture of the nucleation mechanism, quantifying and rationalizing the appearance of stacking disorder based on the different properties of hexagonal and cubic ice. In several cases, our findings corroborate the results of previous studies, based on different potentials [38] or sampling algorithms [34, 42, 40], and therefore contribute to enlarge and put on solid numerical bases the current understanding of homogeneous ice nucleation mechanism. In particular, we directly observed the spontaneous evolution both of a hexagonal and a cubic crystal seed, and we drew statistically reliable conclusions thanks to the repetition of each type of simulation in 15 independent runs. The insight we obtained includes a two-step mechanism for the aggregation of new hexagonal ice molecules to the critical nucleus, compared to a one-step process for the addition of cubic ice molecules. Clearly, our results are in stark contrast with the idealized picture of classical nucleation theory,

assuming a spherical nucleus with the crystalline structure of the thermodynamically stable polymorph (hexagonal ice), and confirm the need of atomistic molecular dynamics simulations with accurate potentials to obtain qualitative and quantitative information on the nucleation process.

One of the outcomes of our study consists in the first quantitative assessment and validation of the PIV metric by means of the rigorous likelihood optimization technique of Ref. [88] : tested against extensive committor data, this coordinate outperforms the whole set of 26 nucleation coordinates analyzed in Ref. [38], encompassing varied information like nucleus size, shape and energy. We remark that our definition – differently from most other definitions found in the ice nucleation literature – is very general, requiring only to define reference initial and end states, hence it could find effective application beyond water, to study the nucleation of other materials as well as a range of other transformation processes in condensed matter. Ongoing applications in our group include salt precipitation from solutions and nanoparticle transformations.

Finally, the large data set of transition pathways obtained in this study will constitute the basis of a new research project, aiming at accurate nucleation rates through a Bayesian reconstruction of a diffusive Markov model starting from path sampling information, along the lines described in Ref. [43]. This approach will tackle what is arguably the most important question – still open despite all recent theoretical and numerical efforts – in the broad field of homogeneous (but also heterogeneous) crystal nucleation : how to predict accurate nucleation rates for ice and other materials in a direct and reliable way, avoiding the approximations (and numerical pitfalls) of classical nucleation theory.

8 Conclusion

In this thesis, we addressed and shed new light on several open questions about water properties in the supercooled regime. We achieved this result by successfully applying general computer simulation methods, that were not devised for the specific case of water, but as general tools to study the broad domain of structural transformations of condensed matter.

Our computational strategy is based on the adoption of realistic and accurate interatomic potentials for water (the TIP4P family) and on flexible generalized coordinates to describe complex transformation processes. Such coordinates are constructed from a simple definition of state vector, the so-called Permutation Invariant Vector (PIV), that is able to store detailed information about the topology of the network of interatomic connections. The latter is defined in a broad sense, extending beyond covalent- and non-covalent bonds until a pre-defined distance range. From this viewpoint, a significant structural transformation of matter (i.e., not a simple deformation preserving the interatomic connections) corresponds to a modification of the network, thus providing a unified framework independent from the particular nature of the material under study. The importance of effective reaction coordinates cannot be underestimated as they are the most crucial ingredient of enhanced sampling algorithms, without which it is unfeasible to reconstruct mechanisms, free-energy landscapes and kinetic rates of rare events like phase transitions.

We employed a rich toolbox of state-of-the-art enhanced sampling techniques, ranging from metadynamics to umbrella sampling to transition path sampling, in the attempt to extract for each scientific problem the most reliable and statistically precise results at an affordable computational cost. Combined with massive computer resources, these simulation methods allowed us to obtain new insight on two of the most challenging problems in the simulation community, at the center of many research efforts : the liquid-liquid transition and the homogeneous nucleation mechanism of supercooled water.

For the first problem, presented in chapter 6, we used extensive molecular dynamics simulations with the TIP4P/2005 model to compute in a systematic way precise free-energy profiles for several conditions of pressure and temperature, deep in the no man's land region. This accomplishment, frequently invoked by the community as a way-out of decades-long controversies [110, 1, 26] but never achieved before, allows us to conclude that there is no free-energy barrier related to a discontinuous, first-order liquid-liquid phase transition – at least down to 155 K, for this accurate model of water and for the system size we considered (800 molecules).

We estimate that future directions of progress in this topic include simulations on larger system, on a longer time scale (especially at lower temperature), and in the thermodynamic region connecting supercooled liquid water with the amorphous forms, that have been extensively studied also from the experimental viewpoint in the last decades. Such efforts, today at the frontier of feasibility in terms of computer resources, could help complete the complex picture of metastable supercooled water. Moreover, further theoretical work will – hopefully – allow to clarify how to reconcile the multiple numerical indications of a second critical point, as put forward in the literature, with our direct observation of barrier-less free-energy landscapes at much lower temperature.

For the second problem, presented in chapter 7, we reconstructed the transition path ensemble for homogeneous ice nucleation from supercooled water using extensive molecular dynamics simulations with the TIP4P/Ice model. For the first time we could apply the expensive but rigorous aimless shooting technique, previously limited to a less realistic coarse-grained model of water, to draw robust conclusions about the structure of critical nuclei and about the mechanism of incorporation of stacking disorder, directly simulating the evolution towards optimal transition states of purely hexagonal and purely cubic nuclei. All these findings are in striking contrast with the hypotheses of classical nucleation theory, and underline once again the need for atomistic simulations to obtain correct qualitative and quantitative information about nucleation processes.

From a methodological viewpoint, our study also allowed to demonstrate the major gain in efficiency represented by a recently developed variant of the aimless shooting algorithm, exploiting the peculiar diffusion properties of the order parameter for nucleation when close to critical size. We believe that this finding will be useful to other researchers studying nucleation of ice or other materials with transition path sampling techniques, greatly accelerating their simulations at the price of a minor modification of the algorithm. An obvious future direction being the study of heterogeneous ice nucleation.

In addition to obtaining detailed information on the ice nucleation mechanism, we performed a quantitative assessment of the quality of the PIV topological metric as reaction coordinate for nucleation : analysis by means of a rigorous likelihood optimization technique based on committor information, indicates that this coordinate outperforms a large set of previously considered coordinates, that tried to capture nucleus size, shape and energy into their definition. This result brings quantitative support to the qualitative observation of the effectiveness of PIV-based generalized coordinates in tracking complex transformation processes in water and other materials, and similar benchmarking is advisable in future projects and for diverse applications.

Finally, our reconstruction of a massive data set of transition pathways for ice nucleation is but the first step towards extracting reliable free-energy landscapes and kinetic rates of the process, what constitutes today a difficult challenge in the simulation community, given the scatter of available predictions. A promising approach in this direction, free from the approximations and pitfalls of formulas derived from classical nucleation theory, appears to be the construction of Markov models (discrete or in the form of a Langevin equation) reproducing in a statistically optimal way the time evolution of the system phase-space point projected on a good reaction coordinate. The transition path ensemble we obtained is the ideal source of data for such objectives.

Références

[1] J. C. Palmer, P. H. Poole, F. Sciortino, and P. G. Debenedetti, "Advances in computational studies of the liquid–liquid transition in water and water-like models," *Chemical reviews*, vol. 118, no. 18, pp. 9129–9151, 2018.

[2] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, "Crystal nucleation in liquids : Open questions and future challenges in molecular dynamics simulations," *Chemical reviews*, vol. 116, no. 12, pp. 7078–7116, 2016.

[3] V. Molinero and E. B. Moore, "Water modeled as an intermediate element between carbon and silicon," *The Journal of Physical Chemistry B*, vol. 113, no. 13, pp. 4008–4016, 2009. PMID : 18956896.

[4] F. H. Stillinger and A. Rahman, "Improved simulation of liquid water by molecular dynamics," *The Journal of Chemical Physics*, vol. 60, no. 4, pp. 1545–1557, 1974.

[5] J. L. Abascal and C. Vega, "A general purpose model for the condensed phases of water : Tip4p/2005," *The Journal of chemical physics*, vol. 123, no. 23, p. 234505, 2005.

[6] J. L. F. Abascal, E. Sanz, R. García Fernández, and C. Vega, "A potential model for the study of ices and amorphous water : Tip4p/ice," *The Journal of Chemical Physics*, vol. 122, no. 23, p. 234511, 2005.

[7] G. A. Gallet and F. Pietrucci, "Structural cluster analysis of chemical reactions in solution," *J. Chem. Phys.*, vol. 139, no. 7, p. 074101, 2013.

[8] 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.

[9] A. H. Nguyen and V. Molinero, "Identification of clathrate hydrates, hexagonal ice, cubic ice, and liquid water in simulations : the chill+ algorithm," *The Journal of Physical Chemistry B*, vol. 119, no. 29, pp. 9369–9376, 2015. PMID : 25389702.

[10] P. L. Geissler, C. Dellago, and D. Chandler, "Kinetic pathways of ion pair dissociation in water," *The Journal of Physical Chemistry B*, vol. 103, no. 18, pp. 3706–3710, 1999.

[11] R. B. Best and G. Hummer, "Reaction coordinates and rates from transition paths," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6732–6737, 2005.

[12] B. Peters and B. L. Trout, "Obtaining reaction coordinates by likelihood maximization," *The Journal of chemical physics*, vol. 125, no. 5, p. 054108, 2006.

[13] B. Peters, G. T. Beckham, and B. L. Trout, "Extensions to the likelihood maximization approach for finding reaction coordinates," *The Journal of chemical physics*, vol. 127, no. 3, p. 034109, 2007.

[14] G. Torrie and J. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation : Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187 – 199, 1977.

[15] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "Transition path sampling : Throwing ropes over rough mountain passes, in the dark," *Annual review of physical chemistry*, vol. 53, no. 1, pp. 291–318, 2002.

[16] O. Mishima, L. D. Calvert, and E. Whalley, "An apparently first-order transition between two amorphous phases of ice induced by pressure," *Nature*, vol. 314, no. 6006, pp. 76–78, 1985.

[17] O. Mishima, L. D. Calvert, and E. Whalley, "'melting ice' i at 77 k and 10 kbar : a new method of making amorphous solids," *Nature*, vol. 310, no. 5976, pp. 393–395, 1984.

[18] J. Finney, D. Bowron, A. Soper, T. Loerting, E. Mayer, and A. Hallbrucker, "Structure of a new dense amorphous ice," *Physical review letters*, vol. 89, no. 20, p. 205503, 2002.

[19] S. Klotz, T. Strässle, R. J. Nelmes, J. S. Loveday, G. Hamel, G. Rousse, B. Canny, J. C. Chervin, and A. M. Saitta, "Nature of the polyamorphic transition in ice under pressure," *Phys. Rev. Lett.*, vol. 94, p. 025506, Jan 2005.

[20] P. H. Poole, F. Sciortino, U. Essmann, and H. E. Stanley, "Phase behaviour of metastable water," *Nature*, vol. 360, no. 6402, pp. 324–328, 1992.

[21] L. E. Bove, F. Pietrucci, A. M. Saitta, S. Klotz, and J. Teixeira, "On the link between polyamorphism and liquid-liquid transition : The case of salty water," *J. Chem. Phys.*, vol. 151, no. 4, p. 044503, 2019.

[22] L. E. Bove, S. Klotz, J. Philippe, and A. M. Saitta, "Pressure-induced polyamorphism in salty water," *Phys. Rev. Lett.*, vol. 106, p. 125701, Mar 2011.

[23] J. C. Palmer, F. Martelli, Y. Liu, R. Car, A. Z. Panagiotopoulos, and P. G. Debenedetti, "Metastable liquid–liquid transition in a molecular model of water," *Nature*, vol. 510, no. 7505, pp. 385–388, 2014.

[24] R. S. Singh, J. W. Biddle, P. G. Debenedetti, and M. A. Anisimov, "Two-state thermodynamics and the possibility of a liquid-liquid phase transition in supercooled tip4p/2005 water," *J. Chem. Phys.*, vol. 144, no. 14, p. 144504, 2016.

[25] J. W. Biddle, R. S. Singh, E. M. Sparano, F. Ricci, M. A. González, C. Valeriani, J. L. Abascal, P. G. Debenedetti, M. A. Anisimov, and F. Caupin, "Two-structure thermodynamics for the tip4p/2005 model of water covering supercooled and deeply stretched regions," *J. Chem. Phys.*, vol. 146, no. 3, p. 034502, 2017.

[26] P. G. Debenedetti, F. Sciortino, and G. H. Zerze, "Second critical point in two realistic models of water," *Science*, vol. 369, no. 6501, pp. 289–292, 2020.

[27] K. Komatsu, S. Machida, F. Noritake, T. Hattori, A. Sano-Furukawa, R. Yamane, K. Yamashita, and H. Kagi, "Ice i c without stacking disorder by evacuating hydrogen from hydrogen hydrate," *Nature communications*, vol. 11, no. 1, pp. 1–5, 2020.

[28] L. Del Rosso, M. Celli, F. Grazzi, M. Catti, T. C. Hansen, A. D. Fortes, and L. Ulivi, "Cubic ice ic without stacking defects obtained from ice xvii," *Nature Materials*, pp. 1–6, 2020.

[29] W. Kuhs, D. Bliss, and J. Finney, "High-resolution neutron powder diffraction study of ice ic," *Le Journal de Physique Colloques*, vol. 48, no. C1, pp. C1–631, 1987.

[30] W. F. Kuhs, C. Sippel, A. Falenty, and T. C. Hansen, "Extent and relevance of stacking disorder in "ice ic"," *Proceedings of the National Academy of Sciences*, vol. 109, no. 52, pp. 21259–21264, 2012.

[31] T. L. Malkin, B. J. Murray, C. G. Salzmann, V. Molinero, S. J. Pickering, and T. F. Whale, "Stacking disorder in ice i," *Physical Chemistry Chemical Physics*, vol. 17, no. 1, pp. 60–76, 2015.

[32] E. Sanz, C. Vega, J. Espinosa, R. Caballero-Bernal, J. Abascal, and C. Valeriani, "Homogeneous ice nucleation at moderate supercooling from molecular simulation," *Journal of the American Chemical Society*, vol. 135, no. 40, pp. 15008–15017, 2013.

[33] J. Espinosa, E. Sanz, C. Valeriani, and C. Vega, "Homogeneous ice nucleation evaluated for several water models," *The Journal of chemical physics*, vol. 141, no. 18, p. 18C529, 2014.

[34] J. R. Espinosa, C. Vega, C. Valeriani, and E. Sanz, "Seeding approach to crystal nucleation," *The Journal of Chemical Physics*, vol. 144, no. 3, p. 034501, 2016.

[35] J. R. Espinosa, C. Vega, C. Valeriani, and E. Sanz, "Seeding approach to crystal nucleation," *The Journal of chemical physics*, vol. 144, no. 3, p. 034501, 2016.

[36] J. Espinosa, C. Navarro, E. Sanz, C. Valeriani, and C. Vega, "On the time required to freeze water," *The Journal of chemical physics*, vol. 145, no. 21, p. 211922, 2016.

[37] A. Haji-Akbari, "Forward-flux sampling with jumpy order parameters," *The Journal of chemical physics*, vol. 149, no. 7, p. 072303, 2018.

[38] L. Lupi, A. Hudait, B. Peters, M. Grünwald, R. G. Mullen, A. H. Nguyen, and V. Molinero, "Role of stacking disorder in ice nucleation," *Nature*, vol. 551, no. 7679, pp. 218–222, 2017.

[39] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12562–12566, 2002.

[40] H. Niu, Y. I. Yang, and M. Parrinello, "Temperature dependence of homogeneous nucleation in ice," *Physical review letters*, vol. 122, no. 24, p. 245501, 2019.

[41] H. Jung, K.-i. Okazaki, and G. Hummer, "Transition path sampling of rare events by shooting from the top," *The Journal of chemical physics*, vol. 147, no. 15, p. 152716, 2017.

[42] A. Haji-Akbari and P. G. Debenedetti, "Direct calculation of ice homogeneous nucleation rate for a molecular model of water," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. 10582–10588, 2015.

[43] M. Innerbichler, G. Menzl, and C. Dellago, "State-dependent diffusion coefficients and free energies for nucleation processes from bayesian trajectory analysis," *Molecular physics*, vol. 116, no. 21-22, pp. 2987–2997, 2018.

[44] F. Caupin, "Escaping the no man's land : Recent experiments on metastable liquid water," *Journal of Non-Crystalline Solids*, vol. 407, pp. 441 – 448, 2015. 7th IDMRCS : Relaxation in Complex Systems.

[45] M. A. Anisimov, M. Duška, F. Caupin, L. E. Amrhein, A. Rosenbaum, and R. J. Sadus, "Thermodynamics of fluid polyamorphism," *Physical Review X*, vol. 8, no. 1, p. 011004, 2018.

[46] F. Pietrucci, "Strategies for the exploration of free energy landscapes : unity in diversity and challenges ahead," *Reviews in Physics*, vol. 2, pp. 32–45, 2017.

[47] 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.

[48] M. F. Chaplin, "Structure and properties of water in its various states," *Encyclopedia of Water : Science, Technology, and Society*, pp. 1–19, 2019.

[49] P. A. Kollman and L. C. Allen, "Theory of the hydrogen bond," *Chemical Reviews*, vol. 72, no. 3, pp. 283–303, 1972.

[50] T. Meier, S. Petitgirard, S. Khandarkhaeva, and L. Dubrovinsky, "Observation of nuclear quantum effects and hydrogen bond symmetrisation in high pressure ice," *Nature communications*, vol. 9, no. 1, pp. 1–7, 2018.

[51] A. Nilsson and L. Pettersson, "Perspective on the structure of liquid water," *Chemical Physics*, vol. 389, no. 1, pp. 1–34, 2011.

[52] L. P. Singh, B. Issenmann, and F. Caupin, "Pressure dependence of viscosity in supercooled water and a unified approach for thermodynamic and dynamic anomalies of water," *Proceedings of the National Academy of Sciences*, 2017.

[53] F. Perakis, K. Amann-Winkel, F. Lehmkühler, M. Sprung, D. Mariedahl, J. A. Sellberg, H. Pathak, A. Späh, F. Cavalca, D. Schlesinger, A. Ricci, A. Jain, B. Massani, F. Aubree, C. J. Benmore, T. Loerting, G. Grübel, L. G. M. Pettersson, and A. Nilsson, "Diffusive dynamics during the high-to-low density transition in amorphous ice," *Proceedings of the National Academy of Sciences*, vol. 114, no. 31, pp. 8193–8198, 2017.

[54] C. Lin, J. S. Smith, S. V. Sinogeikin, and G. Shen, "Experimental evidence of low-density liquid water upon rapid decompression," *Proceedings of the National Academy of Sciences*, vol. 115, no. 9, pp. 2010–2015, 2018.

[55] H. Pathak, A. Späh, K. H. Kim, I. Tsironi, D. Mariedahl, M. Blanco, S. Huotari, V. Honkimäki, and A. Nilsson, "Intermediate range o–o correlations in supercooled water down to 235 k," *J. Chem. Phys.*, vol. 150, no. 22, p. 224506, 2019.

[56] S. Overduin and G. Patey, "An analysis of fluctuations in supercooled tip4p/2005 water," *J. Chem. Phys.*, vol. 138, no. 18, p. 184502, 2013.

[57] S. Overduin and G. Patey, "Fluctuations and local ice structure in model supercooled water," *J. Chem. Phys.*, vol. 143, no. 9, p. 094504, 2015.

[58] H. J. Berendsen and W. F. Van Gunsteren, "Practical algorithms for dynamic simulations," *Molecular-dynamics simulation of statistical-mechanical systems*, pp. 43–65, 1986.

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

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

[61] B. Cooke and S. C. Schmidler, "Preserving the boltzmann ensemble in replica-exchange molecular dynamics," *The Journal of chemical physics*, vol. 129, no. 16, p. 164112, 2008.

[62] M. R. Shirts, "Simple quantitative tests to validate sampling from thermodynamic ensembles," *Journal of chemical theory and computation*, vol. 9, no. 2, pp. 909–926, 2013.

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

[64] S. U. I. NOSÉ, "A molecular dynamics method for simulations in the canonical ensemble," *Molecular Physics*, vol. 100, no. 1, pp. 191–198, 2002.

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

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

[67] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals : A new molecular dynamics method," *Journal of Applied physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[68] S. Nosé and M. Klein, "Constant pressure molecular dynamics for molecular systems," *Molecular Physics*, vol. 50, no. 5, pp. 1055–1076, 1983.

[69] P. P. Ewald, "Die berechnung optischer und elektrostatischer gitterpotentiale," *Annalen der physik*, vol. 369, no. 3, pp. 253–287, 1921.

[70] T. Darden, D. York, and L. Pedersen, "Particle mesh ewald : An $n \cdot \log(n)$ method for ewald sums in large systems," *The Journal of chemical physics*, vol. 98, no. 12, pp. 10089–10092, 1993.

[71] F. Smallenburg, P. H. Poole, and F. Sciortino, "Phase diagram of the st2 model of water," *Molecular Physics*, vol. 113, no. 17-18, pp. 2791–2798, 2015.

[72] C. Vega, J. L. F. Abascal, M. M. Conde, and J. L. Aragones, "What ice can teach us about water interactions : a critical comparison of the performance of different water models," *Faraday Discuss.*, vol. 141, pp. 251–276, 2009.

[73] D. Corradini, M. Rovere, and P. Gallo, "A route to explain water anomalies from results on an aqueous solution of salt," *J. Chem. Phys.*, vol. 132, no. 13, p. 134508, 2010.

[74] M. Conde, M. Gonzalez, J. Abascal, and C. Vega, "Determining the phase diagram of water from direct coexistence simulations : The phase diagram of the tip4p/2005 model revisited," *J. Chem. Phys.*, vol. 139, no. 15, p. 154505, 2013.

[75] M. Conde, M. Rovere, and P. Gallo, "High precision determination of the melting points of water tip4p/2005 and water tip4p/ice models by the direct coexistence technique," *J. Chem. Phys.*, vol. 147, no. 24, p. 244506, 2017.

[76] V. Holten, D. T. Limmer, V. Molinero, and M. A. Anisimov, "Nature of the anomalies in the supercooled liquid state of the mw model of water," *J. Chem. Phys.*, vol. 138, no. 17, p. 174501, 2013.

[77] F. Romano, J. Russo, and H. Tanaka, "Novel stable crystalline phase for the stillinger-weber potential," *Physical Review B*, vol. 90, no. 1, p. 014204, 2014.

[78] T. Li, D. Donadio, G. Russo, and G. Galli, "Homogeneous ice nucleation from supercooled water," *Physical Chemistry Chemical Physics*, vol. 13, no. 44, pp. 19807–19813, 2011.

[79] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "Plumed 2 : New feathers for an old bird," *Computer Phys. Commun.*, vol. 185, no. 2, pp. 604–613, 2014.

[80] A. Stukowski, "Visualization and analysis of atomistic simulation data with ovito-the open visualization tool," *Modelling And Simulation In Materials Science And Engineering*, vol. 18, Jan. 2010.

[81] S. Harrington, R. Zhang, P. H. Poole, F. Sciortino, and H. E. Stanley, "Liquid-liquid phase transition : Evidence from simulations," *Phys. Rev. Lett.*, vol. 78, no. 12, p. 2409, 1997.

[82] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," *Physical Review B*, vol. 28, no. 2, p. 784, 1983.

[83] W. Lechner and C. Dellago, "Accurate determination of crystal structures based on averaged local bond order parameters," *The Journal of Chemical Physics*, vol. 129, no. 11, p. 114707, 2008.

[84] F. Pietrucci and R. Martoňák, "Systematic comparison of crystalline and amorphous phases : Charting the landscape of water structures and transformations," *J. Chem. Phys.*, vol. 142, no. 10, p. 104704, 2015.

[85] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From a to b in free energy space," *The Journal of Chemical Physics*, vol. 126, no. 5, p. 054103, 2007.

[86] S. Jungblut and C. Dellago, "Pathways to self-organization : Crystallization via nucleation and growth," *The European Physical Journal E*, vol. 39, no. 8, p. 77, 2016.

[87] M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering mountain passes via torchlight : Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions," *Annual review of physical chemistry*, vol. 64, pp. 295–316, 2013.

[88] B. Peters, "Reaction coordinates and mechanistic hypothesis tests," *Annual review of physical chemistry*, vol. 67, pp. 669–690, 2016.

[89] P. V. Banushkina and S. V. Krivov, "Optimal reaction coordinates," *Wiley Interdisciplinary Reviews : Computational Molecular Science*, vol. 6, no. 6, pp. 748–763, 2016.

[90] M. Fitzner, G. C. Sosso, F. Pietrucci, S. Pipolo, and A. Michaelides, "Pre-critical fluctuations and what they disclose about heterogeneous crystal nucleation," *Nat. Commun.*, vol. 8, no. 1, pp. 1–7, 2017.

[91] 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.

[92] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *J. Chem Phys.*, vol. 129, no. 12, p. 124105, 2008.

[93] Y. Sugita, A. Kitao, and Y. Okamoto, "Multidimensional replica-exchange method for free-energy calculations," *The Journal of Chemical Physics*, vol. 113, no. 15, pp. 6042–6051, 2000.

[94] 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.

[95] 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.

[96] G. Bussi and A. Laio, "Using metadynamics to explore complex free-energy landscapes," *Nature Reviews Physics*, pp. 1–13, 2020.

[97] G. Bussi, A. Laio, and M. Parrinello, "Equilibrium free energies from nonequilibrium metadynamics," *Physical review letters*, vol. 96, no. 9, p. 090601, 2006.

[98] A. Laio and F. L. Gervasio, "Metadynamics : a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Reports on Progress in Physics*, vol. 71, no. 12, p. 126601, 2008.

[99] A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics," *Wiley Interdisciplinary Reviews : Computational Molecular Science*, vol. 1, no. 5, pp. 826–843, 2011.

[100] E. Vanden-Eijnden *et al.*, "Transition-path theory and path-finding algorithms for the study of rare events.," *Annual review of physical chemistry*, vol. 61, pp. 391–420, 2010.

[101] P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Illustration of transition path theory on a collection of simple examples," *The Journal of chemical physics*, vol. 125, no. 8, p. 084110, 2006.

[102] X.-M. Bai and M. Li, "Calculation of solid-liquid interfacial free energy : A classical nucleation theory based approach," *The Journal of Chemical Physics*, vol. 124, no. 12, p. 124707, 2006.

[103] G. T. Beckham, B. Peters, and B. L. Trout, "Evidence for a size dependent nucleation mechanism in solid state polymorph transformations," *The Journal of Physical Chemistry B*, vol. 112, no. 25, pp. 7460–7466, 2008.

[104] G. T. Beckham and B. Peters, "Optimizing nucleus size metrics for liquid–solid nucleation from transition paths of near-nanosecond duration," *The Journal of Physical Chemistry Letters*, vol. 2, no. 10, pp. 1133–1138, 2011.

[105] M. Marriott, L. Lupi, A. Kumar, and V. Molinero, "Following the nucleation pathway from disordered liquid to gyroid mesophase," *The Journal of chemical physics*, vol. 150, no. 16, p. 164902, 2019.

[106] K.-i. Okazaki, D. Wöhlert, J. Warnau, H. Jung, Ö. Yildiz, W. Kühlbrandt, and G. Hummer, "Mechanism of electroneutral sodium/proton antiporter from transition-path shooting," *bioRxiv*, p. 538777, 2019.

[107] T. Bartels-Rausch, V. Bergeron, J. H. Cartwright, R. Escribano, J. L. Finney, H. Grothe, P. J. Gutiérrez, J. Haapala, W. F. Kuhs, J. B. Pettersson, *et al.*, "Ice structures, patterns, and processes : A view across the icefields," *Reviews of Modern Physics*, vol. 84, no. 2, p. 885, 2012.

[108] O. Mishima and H. E. Stanley, "The relationship between liquid, supercooled and glassy water," *Nature*, vol. 396, no. 6709, pp. 329–335, 1998.

[109] K. Himoto, M. Matsumoto, and H. Tanaka, "Yet another criticality of water," *Physical Chemistry Chemical Physics*, vol. 16, no. 11, pp. 5081–5087, 2014.

[110] K. Amann-Winkel, R. Böhmer, F. Fujara, C. Gainaru, B. Geil, and T. Loerting, "Colloquium : Water's controversial glass transitions," *Reviews of Modern Physics*, vol. 88, no. 1, p. 011002, 2016.

[111] P. Gallo, K. Amann-Winkel, C. A. Angell, M. A. Anisimov, F. Caupin, C. Chakravarty, E. Lascaris, T. Loerting, A. Z. Panagiotopoulos, J. Russo, *et al.*, "Water : A tale of two liquids," *Chemical reviews*, vol. 116, no. 13, pp. 7463–7500, 2016.

[112] H. Tanaka, "Simple physical model of liquid water," *J. Chem. Phys.*, vol. 112, no. 2, pp. 799–809, 2000.

[113] V. Holten and M. A. Anisimov, "Entropy-driven liquid-liquid separation in supercooled water," *Scientific Report*, vol. 713, no. 2, 2012.

[114] J. Russo and H. Tanaka, "Understanding water's anomalies with locally favoured structures," *Nature Communication*, vol. 3556, no. 5, 2014.

[115] R. Esposito, F. Saija, A. Marco Saitta, and P. V. Giaquinta, "Entropy-based measure of structural order in water," *Phys. Rev. E*, vol. 73, p. 040502, Apr 2006.

[116] R. J. Nelmes, J. S. Loveday, T. Strässle, C. L. Bull, M. Guthrie, G. Hamel, and S. Klotz, "Annealed high-density amorphous ice under pressure," *Nature Physics*, vol. 2, pp. 414–418, 2006.

[117] A. M. Saitta, T. Strässle, and S. Klotz, "Structural properties of the amorphous ices : An analysis in terms of distance-ranked neighbors and angular correlations," *J. Phys. Chem. B*, vol. 110, no. 8, pp. 3595–3603, 2006. PMID : 16494415.

[118] P. H. Handle and T. Loerting, "Experimental study of the polyamorphism of water. i. the isobaric transitions from amorphous ices to lda at 4 mpa," *J. Chem. Phys.*, vol. 148, no. 12, p. 124508, 2018.

[119] G. Shen, J. S. Smith, and C. Kenney-Benson, "Nature of polyamorphic transformations in h_2O under isothermal compression and decompression," *Phys. Rev. Materials*, vol. 3, p. 073404, Jul 2019.

[120] G. Pallares, M. El Mekki Azouzi, M. A. González, J. L. Aragones, J. L. F. Abascal, C. Valeriani, and F. Caupin, "Anomalies in bulk supercooled water at negative pressure," *Proceedings of the National Academy of Sciences*, vol. 111, no. 22, pp. 7936–7941, 2014.

[121] Y. Liu, J. C. Palmer, A. Z. Panagiotopoulos, and P. G. Debenedetti, "Liquid-liquid transition in st2 water," *J. Chem. Phys.*, vol. 137, no. 21, p. 214505, 2012.

[122] P. H. Poole, R. K. Bowles, I. Saika-Voivod, and F. Sciortino, "Free energy surface of st2 water near the liquid-liquid phase transition," *J. Chem. Phys.*, vol. 138, no. 3, p. 034505, 2013.

[123] D. T. Limmer and D. Chandler, "The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water," *J. Chem. Phys.*, vol. 135, no. 13, p. 134503, 2011.

[124] D. T. Limmer and D. Chandler, "The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. ii," *J. Chem. Phys.*, vol. 138, no. 21, p. 214504, 2013.

[125] J. C. Palmer, A. Haji-Akbari, R. S. Singh, F. Martelli, R. Car, A. Z. Panagiotopoulos, and P. G. Debenedetti, "Comment on "the putative liquid-liquid transition is a liquid-solid transition in atomistic models of water"[i and ii : J. chem. phys. 135, 134503 (2011) ; j. chem. phys. 138, 214504 (2013)]," *J. Chem. Phys.*, vol. 148, no. 13, p. 137101, 2018.

[126] I. Brovchenko, A. Geiger, and A. Oleinikova, "Liquid-liquid phase transitions in supercooled water studied by computer simulations of various water models," *J. Chem. Phys.*, vol. 123, no. 4, p. 044515, 2005.

[127] D. J. Huggins, "Correlations in liquid water for the tip3p-ewald, tip4p-2005, tip5p-ewald, and swm4-ndp models," *J. Chem. Phys.*, vol. 136, no. 6, p. 064518, 2012.

[128] J. L. Abascal and C. Vega, "Widom line and the liquid–liquid critical point for the tip4p/2005 water model," *J. Chem. Phys.*, vol. 133, no. 23, p. 234502, 2010.

[129] D. T. Limmer and D. Chandler, "Time scales of supercooled water and implications for reversible polyamorphism," *Molecular Physics*, vol. 113, no. 17-18, pp. 2799–2804, 2015.

[130] J. C. Palmer, R. S. Singh, R. Chen, F. Martelli, and P. G. Debenedetti, "Density and bond-orientational relaxations in supercooled water," *Molecular Physics*, vol. 114, no. 18, pp. 2580–2585, 2016.

[131] T. Yagasaki, M. Matsumoto, and H. Tanaka, "Spontaneous liquid-liquid phase separation of water," *Physical Review E*, vol. 89, no. 2, p. 020301, 2014.

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

[133] 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.

[134] J. Goings, "python implementation of binless wham, https ://github.com/jjgoings/wham," 2018.

[135] A. Grossfield, "Wham : the weighted histogram analysis method, version 2.0.9, `http://membrane.urmc.rochester.edu/wordpress/?page_id=126`," 2018.

[136] J. Wong, D. A. Jahn, and N. Giovambattista, "Pressure-induced transformations in glassy water : A computer simulation study using the tip4p/2005 model," *J. Chem. Phys.*, vol. 143, no. 7, p. 074501, 2015.

[137] A. Humphrey W., Dalke and K. Schulten, "Vmd - visual molecular dynamics," *J. Molec. Graphics*, vol. 14.1, pp. 33–38, 1996.

[138] A. Humphrey W., Dalke and K. Schulten, "Vmd - quicksurf," 2020.

[139] F. Martelli, "Unravelling the contribution of local structures to the anomalies of water : The synergistic action of several factors," *J. Chem. Phys.*, vol. 150, no. 9, p. 094506, 2019.

[140] J. Guo, R. S. Singh, and J. C. Palmer, "Anomalous scattering in supercooled st2 water," *Molecular Physics*, vol. 116, no. 15-16, pp. 1953–1964, 2018.

[141] H. Vehkamäki, *Classical nucleation theory in multicomponent systems*. Springer Science & Business Media, 2006.

[142] V. I. Kalikmanov, "Classical nucleation theory," in *Nucleation theory*, pp. 17–41, Springer, 2013.

[143] R. P. Sear, "Nucleation : theory and applications to protein solutions and colloidal suspensions," *Journal of Physics : Condensed Matter*, vol. 19, no. 3, p. 033101, 2007.

[144] R. P. Sear, "The non-classical nucleation of crystals : microscopic mechanisms and applications to molecular crystals, ice and calcium carbonate," *International Materials Reviews*, vol. 57, no. 6, pp. 328–356, 2012.

[145] B. J. Murray, C. G. Salzmann, A. J. Heymsfield, S. Dobbie, R. R. Neely III, and C. J. Cox, "Trigonal ice crystals in earth's atmosphere," *Bulletin of the American Meteorological Society*, vol. 96, no. 9, pp. 1519–1531, 2015.

[146] T. L. Malkin, B. J. Murray, A. V. Brukhno, J. Anwar, and C. G. Salzmann, "Structure of ice crystallized from supercooled water," *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1041–1045, 2012.

[147] B. J. Murray and A. K. Bertram, "Formation and stability of cubic ice in water droplets," *Physical Chemistry Chemical Physics*, vol. 8, no. 1, pp. 186–192, 2006.

[148] T. Hansen, M. Koza, and W. Kuhs, "Formation and annealing of cubic ice : I. modelling of stacking faults," *Journal of Physics : Condensed Matter*, vol. 20, no. 28, p. 285104, 2008.

[149] A. J. Amaya, H. Pathak, V. P. Modak, H. Laksmono, N. D. Loh, J. A. Sellberg, R. G. Sierra, T. A. McQueen, M. J. Hayes, G. J. Williams, *et al.*, "How cubic can ice be ?," *The Journal of Physical Chemistry Letters*, vol. 8, no. 14, pp. 3216–3222, 2017.

[150] E. B. Moore and V. Molinero, "Structural transformation in supercooled water controls the crystallization rate of ice," *Nature*, vol. 479, no. 7374, pp. 506–508, 2011.

[151] A. Hudait, S. Qiu, L. Lupi, and V. Molinero, "Free energy contributions and structural characterization of stacking disordered ices," *Physical Chemistry Chemical Physics*, vol. 18, no. 14, pp. 9544–9553, 2016.

[152] B. Murray, S. Broadley, T. Wilson, S. Bull, R. Wills, H. Christenson, and E. Murray, "Kinetics of the homogeneous freezing of water," *Physical Chemistry Chemical Physics*, vol. 12, no. 35, pp. 10380–10387, 2010.

[153] R. J. Allen, C. Valeriani, and P. R. ten Wolde, "Forward flux sampling for rare event simulations," *Journal of physics : Condensed matter*, vol. 21, no. 46, p. 463102, 2009.

[154] S. Hussain and A. Haji-Akbari, "Studying rare events using forward-flux sampling : Recent breakthroughs and future outlook," *The Journal of Chemical Physics*, vol. 152, no. 6, p. 060901, 2020.

[155] E. Giuffré, S. Prestipino, F. Saija, A. M. Saitta, and P. V. Giaquinta, "Entropy from correlations in tip4p water," *Journal of chemical theory and computation*, vol. 6, no. 3, pp. 625–636, 2010.

[156] J. Grabowska, "Why is the cubic structure preferred in newly formed ice ?," *Phys. Chem. Chem. Phys.*, vol. 21, no. 33, pp. 18043–18047, 2019.