Machine learning methods for computational studies in origins of life

Timothée Devergne

Université Sorbonne Université

École Doctorale 397:
Physique et Chimie des Matériaux

Machine learning methods for computational studies in origins of life

Thèse de Doctorat de Physique

presentée par

Timothée Devergne

Pour l'obtention du grade de:

Docteur de Sorbonne Université

Dirigée par

A. Marco Saitta & Fabio Pietrucci

Soutenue le 18 septembre 2023, devant un jury composé de :

President du jury:	Mathieu Salanne	PHENIX - Sorbonne Université
Rapportrices:	Roberta Poloni	SIMAP - Université Grenoble Alpes
	Helena Zapolsky	GPM - Université Rouen Normandie
Directeurs:	A. Marco Saitta	IMPMC - Sorbonne Université
	Fabio Pietrucci	IMPMC - Sorbonne Université
Examinateurs:	Gabriel Stoltz	CERMICS - Ecole des Ponts
	Ambroise Van Roekeghem	CEA Grenoble

A Patrice et Timy

Remerciements

In order to be understandable by everyone, some parts of the acknowledgement section will be written in English while other in French.

First, I would like to thank all the members of the jury to have accepted to review this thesis and to judge my defense, especially Roberta Poloni and Helena Zapolski who accepted to review this thesis.

Merci à Marco et Fabio de m'avoir encadré pendant trois ans en me donnant des conseils autant scientifiques que personnels. La thèse est comme une réaction chimique qu'on étudie : c'est une montagne à escalader et sans guide pour passer cette barrière, on reste dans le bassin des réactifs.

Lunch and coffee breaks are also an important part of the life of the PhD student. They would not have been that enjoyable without the computational part of the physix team, Sonia, Line, David, Leon, Matteo, Flavio, Karen, Arthur, Hadrien and Zach who became my friends. I hope CROUS bashing will become an olympic discipline so that we can win a medal.

After a big week of work, a beer or two are always welcome to laugh and discuss. Thus, I would address a big thanks also to "the team" Federica, Miha, Matteo Karen for all the supports, beers, and laughs that were always welcome.

Merci à Arthur de m'avoir aidé sur ce manuscript en relisant certains chapitres, pour les séances d'escalade et aussi de m'avoir conseillé de faire un post-doc à mon début de troisième année.

Un simple merci ne suffirait pas pour Mathieu, qui a pris le temps de relire ce manuscript, qui m'a donné de nombreux conseils tout au long de ces trois ans de travail. J'en profite aussi pour remercier Julien Heu de m'avoir accompagné pendant mon stage de master et de m'avoir aidé sur beaucoup de concepts de machine learning.

Comment parler de ces trois années de vie à Paris sans mentionner le tournoi hebdomadaire de Mario Kart qui permettait de sortir un petit peu la tête du guidon, pour ça, merci à Jade, Hayat, Léa, Vincent, Allan, Clément, Cédric.

De même, ce n'est que leur rendre peu de justice que de leur adresser un paragraphe

ici, le duo de légende Jade et Hayat avec qui j'ai passé beaucoup de moments de complicité, de séances de cinéma (bonnes et moins bonnes) durant trois ans.

Merci à Xavier, on a traversé nos études ensemble, on va les finir ensemble, merci d'avoir partagé tous ces instants. Au moment où j'écris ces lignes tu grimpes tranquillement en Grèce, mais prépare toi, c'est bientôt ton tour.

Merci à mes parents Patrick et Odile de m'avoir soutenu pendant ces huit années d'études. Ces huit années ont été longues et parfois compliquées, merci d'avoir toujours supporté mes choix, de m'avoir aidé. Et surtout merci de m'avoir supporté parler de physique pendant les repas. Et bien sûr, au delà des études, merci de m'avoir éduqué comme vous l'avez fait.

Merci à mes soeurs Cécile et Soizic de m'avoir montré l'exemple en faisant de longues études. Merci pour tous les moments de rire que nous avons passés tous les trois. Mais s'il vous plaît ne racontez rien d'embarrassant le jour de la soutenance !

Merci plus généralement à ma famille, aux cousins cousines, Maryline, Anne d'avoir partagé quelques moments en dehors de la physique et de la thèse.

Enfin Louison, il est difficile de résumer en quelques mots à quel point tu m'as aidé et de te remercier à la hauteur de cela. Merci d'accepter de partager ta vie avec moi, pensons au futur maintenant.

J'aimerais terminer ces remerciements en adressant un petit mot à Rémie ma grand-mère qui quand j'ai décidé de faire de la physique m'a dit de faire attention car "on devient fou en faisant ça", j'espère ne pas l'être devenu. Cette thèse est dédiée à mes oncles, Patrice et Timy, qui ont tous les deux joué un grand rôle dans mon éducation, Patrice que j'avais traîné pendant des vacances à Paris dans des expositions scientifiques à la cité des sciences et à la fondation Cartier. Timy qui m'a acheté mon premier numero de Sciences et Vie junior.

Résumé

La chimie prébiotique consiste en l'étude des réactions chimiques aux origines de la vie sur Terre. C'est un très vaste sujet qui mobilise plusieurs domaines scientifiques dont la physique numérique. En effet, des simulations de dynamique moléculaire de haute précision peuvent être menées pour tester l'influence de différents environnements plausibles sur la synthèse de molécules : ce composant a-t-il pu apparaître dans le milieu interstellaire ? Sa formation est-elle favorisée par la présence de surfaces minérales ? Elles peuvent aussi être utilisées pour identifier des intermédiaires réactionnels trop peu stables pour être observés expérimentalement et mieux comprendre les mécanismes de formation. Pour cela, des méthodes d'échantillonnage avancé (EA) comme la metadynamique ou l'umbrella sampling sont utilisées pour explorer et échantillonner l'espace chimique. Ces méthodes peuvent être utilisées par exemple pour étudier la synthèse des acides aminés, qui constituent les briques de base des protéines, des molécules clés pour le vivant. Cela a été fait par Magrino et al., avec l'étude de la synthèse par voie de Strecker de la glycine, l'acide aminé le plus simple, en milieu aqueux. Cela a permis d'identifier tous les intermédiaires réactionnels présents dans cette voie de synthèse et de caractériser leur stabilité relative. Cependant, ces simulations dites ab initio qui prennent en compte les degrés de liberté électroniques ont un coût de calcul élevé et seuls de petits systèmes de l'ordre de la centaine d'atomes peuvent être étudiés. Pour remédier à ce problème, des méthodes d'apprentissage automatique (AA) qui permettent de réduire ce temps de calcul ont été mises en places pour des systèmes à l'équilibre. Peu d'études ont proposé des méthodes d'AA s'appliquant à des événements réactifs qui nécessitent un modèle précis sur l'entièreté de l'espace chimique en conjonction avec l'usage de méthodes d'EA. Dans un premier temps, nous nous appuyons sur les données existantes de la première étape de la synthèse prébiotique de Strecker de la glycine pour développer une méthode d'entraînement de modèles d'AA pour l'étude de réactions chimiques en solution. Nous commençons par entraîner un ensemble de modèles, appelé comité, avec le même ensemble d'entraînement, mais des conditions initiales différentes. Au cours d'une simulation, nous pouvons suivre l'évolution temporelle de la différence de prédictions des forces au sein du modèle et lorsque le système se trouve en dehors de la zone d'entraînement du modèle, nous constatons que cette différence augmente, ce qui nous permet de quantifier la qualité de la prédiction et définir un temps de simulation pendant lequel le modèle se comporte comme une simulation ab initio. Grâce à cela, nous pouvons cibler dans l'espace chimique quelles nouvelles données utiliser pour entraîner un modèle plus performant. En utilisant cette méthode, nous parvenons à obtenir des données proches des données ab initio. Nous appliquons ensuite cette nouvelle méthode à un chemin de synthèse prébiotique de la glycine en milieu aqueux différent de celui de Strecker. Ce chemin n'avait jamais été exploré auparavant. Cependant, la méthode développée précédemment nécessite une connaissance préalable du mécanisme de transition. Dans la deuxième partie de cette thèse, nous utilisons des trajectoires ab initio d'échantillonnage de chemin de transition qui sont des trajectoires démarrant de l'état de transition vers les bassins d'équilibre qui couvrent l'entièreté de l'espace chimique et qui ne demandent pas une caractérisation préalable du mécanisme. Ces trajectoires sont utilisées pour entraîner un modèle qui peut être utilisé pour récupérer les données thermodynamiques, mais aussi cinétiques d'une réaction avec une qualité ab initio pour un moindre coût.

Abstract

Prebiotic chemistry is the study of chemical reactions at the origins of life on Earth. It is a very wide subject that requires the contribution of many scientific fields, including numerical physics. Indeed, highly accurate molecular dynamics simulations are performed to test the influence of different environments on the synthesis of molecules: could this component appear in the interstellar medium? Is its formation impacted by the presence of mineral surfaces? They can also be used to identify intermediates that are too unstable to be observed experimentally and better understand the mechanism of formation.

To do so, enhanced sampling (ES) methods such as metadynamics or umbrella sampling are used to explore and sample the chemical landscape. These methods can be used to study the synthesis of amino acids that are the building blocks of proteins. This was done by Magrino et al., with the study of the Strecker synthesis of glycine, the simplest amino acid, in water. This allowed the identification of all the intermediates and characterization of their stability.

However, these simulations called ab initio, which take into account the electronic degrees of freedom, are computationally expensive, and only small systems of a few hundred atoms can be studied. To solve this problem, machine learning (ML) methods have been put into place that allow the reduction of computational time for equilibrium systems. Only a few ML methods have been suggested to study reactive events because this requires an accurate model across the entire chemical space.

In a first step, we use the existing data from the study of the prebiotic synthesis of glycine to devise a training method for ML models for chemical reactions in solution. We start by training a set of models, called a committee, with the same training set but different initial conditions. During a simulation, we track the evolution of deviation of the prediction of forces, and we see that when the model is out of its training zone, this deviation drastically increases. This allows us to define a time during which the model behaves like an ab initio simulation. Thanks to this, we can target in the chemical space what new data to put in the training set to have a more accurate model. By using this method, we obtained results close to ab initio accuracy.

We then apply this method to a new prebiotic pathway to glycine in water that has never been studied before. However, the method previously developed requires a prior knowledge of the transition mechanism. In the second part of this thesis, we use ab initio transition path sampling trajectories, which are trajectories starting from the transition state and relaxing into the equilibrium basins. They cover all the chemical space and are therefore suitable to train an ML model. By using such a model, we managed not only to recover the thermodynamics of the reaction but also the kinetics. We obtained results close to ab initio accuracy.

Contents

List of Figures

List of Tables

List of Abbreviations

ACSF	Atom Centered Symmetry function
AIMD	*ab initio* molecular dynamics
CV	Collective Variable
DFT	Density Functional Theory
DFTB	Density Functional based Tight Binding
FEP	Free Energy Profile
FES	Free Energy Surface
GGA	Generalized Gradient Approximation
HF	Hartree Fock
ISM	Interstellar Medium
LDA	Local Density Approximation
MD	Molecular dynamics
ML	Machine Learning
MLP	Machine Learning Potential
NN	Neural Network
NNP	Neural Network Potential
PCV	Path Collective Variables
PES	Potential Energy Surface
RC	Reaction Coordinate
RMSE	Root Mean Square Error
TPS	Transition Path Sampling
TST	Transition State Theory
TS	Transition State
WHAM	Weighted Histogram Analysis Method
US	Umbrella Sampling

Résumé en français

La chimie prébiotique consiste en l'étude de réactions ayant eu lieu aux premières heures de la Terre. À cette époque, de petites molécules ont réagi entre elles pour en former de plus grosses, et ce, jusqu'à l'apparition des premières briques du vivant. Il existe beaucoup de scenarii différents pour expliquer ces réactions. En effet, certaines molécules peuvent être arrivées grâce à des impacts de météorites et donc, auraient des origines extra-terrestres. D'autres pourraient être synthétisées dans une "soupe primordiale" pensée par Darwin. Enfin, des éclairs, courants électriques au sein d'une atmosphère composée de gaz simples, auraient pu former les premières biomolécules. Cette dernière possibilité a été testée pour la première fois par S. Miller en laboratoire [1] en faisant passer un courant électrique dans un gaz similaire à ce qui aurait pu être présent sur Terre à cette époque (eau (H_2O), méthane (CH_4), ammoniaque (NH_3), dihydrogène (H_2)). Miller a obtenu un mélange contenant des molécules plus complexes comme la glycine, l'acide aminé le plus simple, montrant ainsi que des composants du vivant auraient pu être formés de cette façon.

En effet, il existe 22 acides aminés différents qui constituent les composants principaux des protéines, qui sont elles-mêmes des molécules nécessaires au vivant : elles participent au métabolisme ainsi qu'à la structuration et à la cohésion des cellules. L'étude de leur structure complexe constitue un pan entier de la physique : la complexité de leur composition fait qu'elles peuvent se replier sur elles-mêmes ou au contraire se déplier. Durant cette thèse, nous nous sommes intéressés à la synthèse prébiotique de la glycine.

Comme exposé précédemment, de nombreuses hypothèses sur la formation des premières biomolécules sur Terre ont été émises. Il est donc difficile pour les chercheuses et les chercheurs de recréer toutes les conditions possibles en laboratoire. Ceci correspond à un premier niveau d'abstraction et pour explorer de nouvelles conditions, nous pouvons passer encore un de ces niveaux en créant une boîte de simulation avec les atomes et molécules voulus. Les déplacements des composants de la boîte sont accessibles en résolvant les équations du mouvement. Cette technique est appelée dynamique moléculaire. Cependant, pour étudier des réactions chimiques, les électrons ont un rôle prépondérant (une réaction chimique peut être vue comme un déplacement d'électrons) et l'équation de Schrödinger qui est l'équation fondamentale de la mécanique quantique, doit être résolue.

Il est possible de la résoudre de façon exacte seulement pour des systèmes comme l'atome d'hydrogène avec un seul électron. Pour obtenir une solution approchée, il existe différentes méthodes dites *ab initio*, car ne nécessitant aucun paramètre extérieur. Parmi toutes ces approximations possibles, nous avons choisi dans ce travail d'utiliser la théorie

de la fonctionnelle de la densité qui permet un bon compromis entre précision et temps de calcul.

Même avec des ressources de calcul infinies, il serait impossible de faire des simulations qui contiennent assez d'information pour avoir une analyse quantitative d'une réaction particulière. En effet, si une boîte de simulation est préparée dans un état et qu'une simulation est lancée depuis celui-ci, le système atteindra un état d'équilibre (soit les réactifs, soit les produits), et ne le quittera plus, car il est trop stable et il est impossible de le quitter. Pour donner une comparaison visuelle, une réaction chimique peut être vue comme le passage d'un col pour aller d'une vallée à une autre (les réactifs et les produits). Notre problème se rapprocherait alors d'une bille à laquelle on aurait donné une petite impulsion en bas du Mont Ventoux : elle parcourrait quelques mètres en direction du sommet, puis redescendrait. Mais il existe des techniques comme la métadynamique [2, 3] pour contourner ce problème. Dans cette méthode, un biais gaussien est régulièrement ajouté dans l'expression de l'énergie potentielle à l'emplacement où se trouve le système, ce qui permet de remplir petit à petit le bassin d'énergie libre et de passer des réactifs vers les produits. Si l'on reprend l'image de la bille au pied du mont Ventoux, cela revient à ajouter régulièrement un petit tas de sable sur la position de la bille jusqu'à ce qu'elle ait atteint le sommet.

La métadynamique nous permet donc d'obtenir une première transition entre les deux vallées. Mais, pour avoir des informations quantitatives sur la réaction, il faut plus de transitions. Pour cela, nous utilisons des simulations d'Umbrella Sampling [4] qui consistent à séparer l'espace chimique en fenêtres. Dans chacune d'elles, une simulation est lancée avec un potentiel quadratique pour restreindre le système autour du centre de la fenêtre. Ceci permet d'avoir des données sur le mécanisme réactionnel tout le long de l'espace chimique. Une fois que toutes les simulations ont été faites dans toutes les fenêtres, les données sont mises ensemble pour obtenir le profil d'énergie libre de la réaction. Grâce à ce profil, il est possible d'étudier la stabilité relative des réactifs et des produits, mais aussi de comparer le mécanisme étudié avec d'autres mécanismes grâce à la barrière d'activation qui donne une information sur la faisabilité de la réaction : un mécanisme dont la barrière d'activation est plus basse arrivera plus rapidement qu'un autre qui a le même point de départ, mais une barrière plus élevée.

Le profil d'énergie libre peut être comparé aux résultats expérimentaux quantitativement à travers la différence d'énergie libre entre les réactifs et les produits. En effet, celle-ci est reliée à la constante d'équilibre de la réaction par une relation logarithmique. Le taux cinétique de la réaction, en d'autres termes la vitesse à laquelle les réactifs se consument, est lui lié à l'exponentielle de l'opposé de la barrière d'énergie libre.

Ces techniques de simulations ont été mises en place par l'équipe dans un protocole méticuleux pendant la thèse de Théo Magrino. Tout d'abord, les réactifs et les produits sont définis. Ensuite, une simulation de métadynamique est lancée pour obtenir une première idée du mécanisme de transition. Après cela, des trajectoires non biaisées sont lancées de points proches du haut de la barrière pour trouver les états de transition.

Il s'agit des configurations pour lesquelles la probabilité de tomber dans le bassin des réactifs est la même que celle de tomber dans le bassin des produits lorsque l'on lance une simulation de ces points. Cela nous permet de mieux comprendre le mécanisme de transition et d'établir une "variable collective", c'est-à-dire la projection des positions des atomes dans une variable qui doit contenir toutes les informations nécessaires à la compréhension du mécanisme. C'est cette quantité qui est ensuite utilisée pour placer les biais quadratiques lors des simulations d'Umbrella Sampling. Dans le protocole établi par Theo Magrino la variable collective est une variable de chemin [5], qui, à partir de configurations de référence sur le chemin réactionnel que nous voulons étudier, permet de situer une configuration le long de ce chemin. Dans notre cas, ces états de référence sont choisis le long des trajectoires non biaisées. Après avoir construit notre variable collective, des simulations d'Umbrella Sampling sont lancées le long de celle-ci (entre 50 et 60 fenêtres). Enfin, le paysage d'énergie libre est obtenu grâce à la méthode d'analyse des histogrammes pondérés (weighted histogram analysis method en anglais).

Ce protocole a permis l'étude de plusieurs réactions [6]. La plus importante et complète d'entre elles est celle du mécanisme de Strecker de formation de la glycine [7]. Cela a permis d'identifier et de caractériser des états intermédiaires qui ne peuvent être vus expérimentalement. Ce travail a mis en évidence la meta-stabilité de l'amino-nitrile associé à la glycine qui est détecté dans le milieu interstellaire contrairement à la glycine.

Cependant, les simulations *ab initio* sont très coûteuses en temps de calcul. Cela limite la taille des systèmes pouvant être étudiés (autour du millier d'atomes et du nanomètre) ainsi que les échelles de temps accessibles (au maximum autour de la nanoseconde). Pour remédier à ce problème, des solutions de machine learning ont été mises en place [8, 9, 10]. Bien qu'elles diffèrent dans leurs applications, le principe de ces méthodes est souvent le même : on dispose des configurations atomiques, souvent générées par dynamique moléculaire *ab initio*, ainsi que les énergies et les forces associées. Le but est alors d'entraîner un modèle capable de prédire ces grandeurs sans avoir à résoudre l'équation de Schrödinger et donc de limiter le temps de calcul. Le modèle obtenu est appelé "potentiel machine learning". Une fois le potentiel entraîné, des simulations de dynamique moléculaire peuvent être lancées avec celui-ci au moyen d'une boîte de simulation plus grande ou pour un temps plus long.

Ces méthodes ont obtenu beaucoup de succès pour des systèmes à l'équilibre ou cristallins [11, 12, 13]. Mais il est beaucoup plus compliqué d'obtenir un potentiel machine learning pour des réactions chimiques. En effet, il faut que celui-ci soit performant sur tout l'espace chimique pour pouvoir échantillonner proprement le mécanisme de transition [14, 15]. C'est ce que nous avons proposé de faire durant cette thèse.

La question qui se pose est donc : Quelles données doit-on utiliser en entraînement d'un modèle machine learning pour pouvoir retrouver le paysage d'énergie libre ? Comment savoir si le modèle de machine learning est fidèle sans avoir à refaire les trajectoires *ab initio*. Nous essayons d'apporter une réponse à ces questions dans l'article publié en août 2022 dans "Journal of chemical theory and computations" [16].

Grâce aux données de l'étude de la synthèse de Strecker de la glycine, nous avons pu établir comment construire efficacement un ensemble d'entraînement. Pour cela, nous avons utilisé la méthode dite des "comités de réseaux de neurones". Pour une étape donnée, nous entraînons quatre réseaux de neurones et nous faisons une simulation avec l'un d'eux. Les autres réseaux sont utilisés pour contrôler la qualité de la simulation : pour chaque pas de temps, on calcule le désaccord entre les membres du comité et s'il est supérieur à un seuil prédéfini, la simulation est arrêtée et le temps de simulation constitue le critère sur lequel nous ajoutons des données dans l'ensemble d'entraînement. En effet, nous effectuons des simulations sur l'ensemble des fenêtres d'Umbrella Sampling, donc sur l'ensemble du chemin réactionnel, ce qui nous permet de voir les zones dans lesquelles le temps de simulation est trop faible et où il faut rajouter des données d'entraînement. Grâce à cette méthode, nous avons pu retrouver les données thermodynamiques de deux étapes de la synthèse de Strecker de la glycine (voir figure 1).

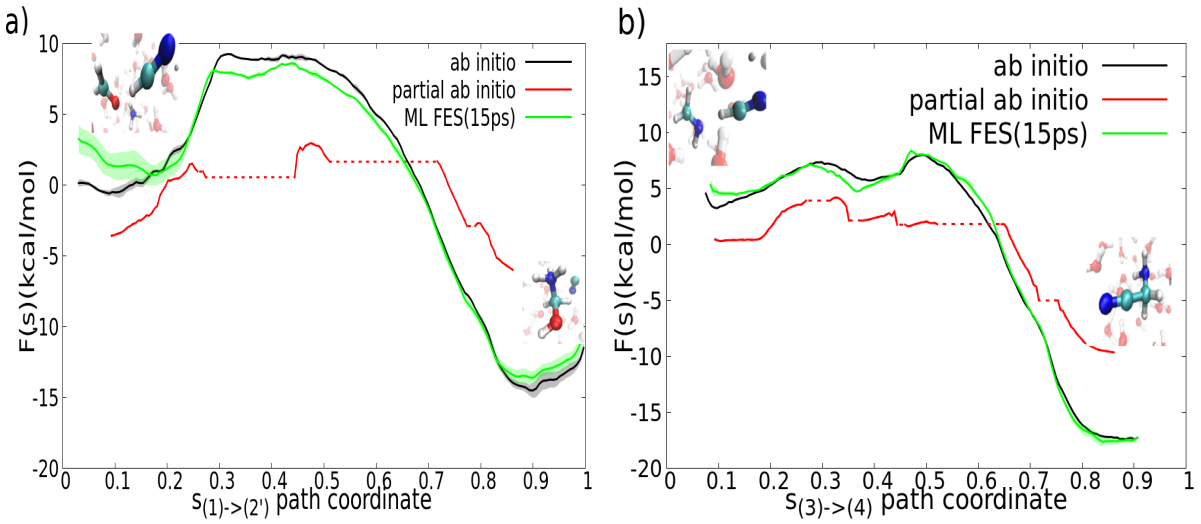

Figure 1: **a)** Énergies libres obtenues *ab initio* (noir), et avec potentiel machine learning (vert) pour la première étape de la synthèse de Strecker de la glycine. **b)** Énergies libres obtenues *ab initio* (noir), et avec potentiel machine learning (vert) pour la quatrième étape de la synthèse de Strecker de la glycine. En rouge, l'énergie libre obtenue avec les seules données présentes dans l'ensemble d'entraînement.

Grâce à cette première étude, nous avons montré qu'il était possible d'établir un profil d'énergie libre à l'aide d'un potentiel machine learning. Nous avons ensuite utilisé cette méthode pour étudier un nouveau mécanisme de formation de la glycine différent de celui de Strecker. En effet, ce dernier considère d'abord l'attaque de l'ammoniaque sur le formaldéhyde. Ici, nous proposons la formation de glycolonitrile par addition de l'ion cyanure sur le formaldéhyde. Le mécanisme fait intervenir un intermédiaire quelque peu exotique, le 2-oxiramine qui est très instable dans l'eau et laisse place à l'acide glycolique. Il a été observé dans des météorites, de même que le glycolonitrile, ce qui vient conforter la plausibilité de notre mécanisme. De plus, les énergies libres que nous avons calculées sont proches de celles relevées expérimentalement comme indiqué dans la figure 2.

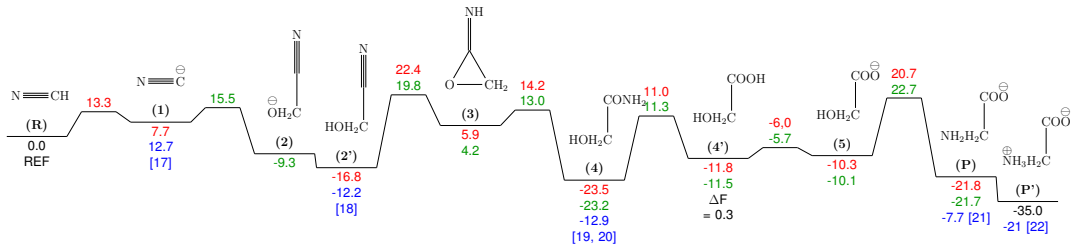

Figure 2: Bilan des énergies libres des différents intermédiaires du mécanisme de formation de la glycine. Les énergies calculées par machine learning sont indiquées en vert, celles calculées *ab initio* en rouge et les données issues de la littérature en bleu.

Contrairement à ce qui avait été fait précédemment pour l'étude du chemin de Strecker, ici, nous ne disposions pas des données de référence avant d'entraîner un potentiel machine learning. Nous avons généré les données d'entraînements au fur et à mesure en suivant la méthode développée dans notre publication [16]. L'accord entre les valeurs *ab initio* et celles machine learning montre une nouvelle fois que cette méthode peut être utilisée pour l'étude de réactions chimiques en solution.

Cependant, l'entraînement de ce potentiel intervient tard dans le protocole d'étude et nécessite la définition préalable d'une variable collective. Des techniques d'exploration de chemin de transition ont récemment été mises en place dans l'équipe, comme le transition path sampling (TPS) [23], pour définir de nouvelles coordonnées de réaction [24]. Le TPS consiste à effectuer un grand nombre de trajectoires démarrant dans la zone de l'état de transition et tombant vers les produits et les réactifs. Cela permet d'échantillonner efficacement tous les chemins permettant de lier les réactifs aux produits. Pour reprendre l'analogie de la bille et de la montagne, faire du TPS reviendrait à vider un sac de billes du haut de la montagne et à regarder le chemin que prend chacune d'elles. Ces trajectoires sont donc composées de configurations le long de toute la transition et donc, peuvent être utilisées pour entraîner un potentiel machine learning. C'est ce que nous avons fait avec les données de TPS obtenues par Théo Magrino et Léon Huet.

Nous avons étudié une réaction très simple, utilisée de nombreuses fois en chimie théorique pour valider des méthodes. Il s'agit de la réaction de substitution du chlorure de méthyle par un ion chlorure dans l'eau. En d'autres termes, la molécule de chlorure de méthyle est remplacée par elle-même, et donc expérimentalement rien n'est observé. Par contre, lors d'une simulation, ce n'est pas le même atome de chlore qui est lié au carbone, et donc, il faut casser une liaison carbone/chlore et en créer une, ce qui implique le passage d'une barrière d'énergie libre. C'est pour cela que des simulations de TPS ont été lancées du haut de la barrière et ont servi à entraîner un potentiel machine learning. Nous l'avons ensuite utilisé pour obtenir le profil d'énergie libre le long d'une variable bien connue : la différence des distances entre le premier atome de chlore et l'atome de carbone et le deuxième atome de chlore et l'atome de carbone. Les résultats obtenus sont

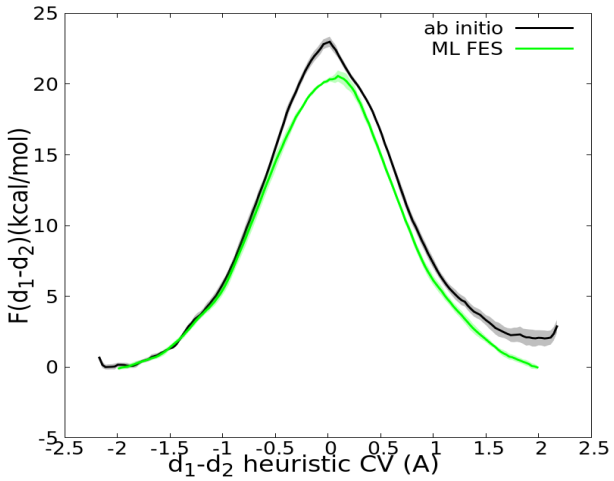

Figure 3: Énergie libre le long d'une variable collective heuristique calculée à partir de distances interatomiques. Les données *ab initio* sont représentées en noir, celles machine learning en vert. Les zones grisées correspondent à l'erreur statistique estimée.

présentés sur la figure 3. Cela montre que le potentiel entraîné est capable de reproduire un profil *ab initio*, et donc, nous avons fait plus de simulations TPS avec ce potentiel. Cela nous a permis de choisir des références pour construire une variable de chemin sur laquelle nous avons aussi fait de l'Umbrella Sampling. Ceci aurait été impossible avec la méthode précédente, car le potentiel est entraîné avec des simulations d'Umbrella Sampling le long d'une variable précise. Nous avons aussi utilisé ces simulations de TPS pour obtenir une estimation précise du taux cinétique grâce au formalisme de Bennett et Chandler [25], ce qui aurait été impossible *ab initio*. En résumé, cette nouvelle méthode nous permet d'obtenir les quantités thermodynamiques et cinétiques permettant de caractériser une transformation physique ou chimique.

Cette nouvelle méthode pourrait être utilisée en chimie prébiotique pour étudier des systèmes plus grands, avec notamment des surfaces minérales agissant comme catalyseurs. Elle pourrait aussi être utilisée dans le cadre de simulations prenant en compte les effets quantiques nucléaires. En effet, ici, nous ne traitons les noyaux que de manière classique et leur comportement est découplé de celui des électrons. Cependant, il existe des systèmes dans lesquels il n'est pas possible de faire cette approximation, notamment les systèmes contenant des atomes légers comme l'atome d'hydrogène. Pour remédier à cela, des simulations coûteuses fonctionnant sur la base de réplicas sont lancées. Cette méthode permettrait de garder une précision *ab initio* en économisant du temps de calcul.

Introduction

In 1972, P.W. Anderson introduced in his now very famous paper "More is different" [26] the fact that even though the fundamental laws of nature are known for the behavior of one or a few bodies, it is not enough to understand the behavior of a large scale system of millions of bodies interacting. This is due to the symmetry breaking and the growing complexity of the system. Many examples of growing complexity can be found in nature, from physics to social sciences: the spiral structure of galaxies from billions of stars [27], the patterns and organization emerging from a flight of birds [28] and the study of the behavior of crowds [29]. All of this cannot be observed by studying only a small number of interacting subjects. However, the most striking emergent phenomenon is probably the emergence of life.

J. H. Conway introduced in 1970 the game of life [30]. On a NxN grid, cells are placed, and their time evolution is defined through four simple rules to decide whether at time t the cell i will die, give birth or do nothing. Although this is a very simple game, many complex emergent behaviors have been observed from it. It has been shown to be Turing complete [31], meaning that it can be used to simulate any other system with any set of rules, including itself. In other words, it is capable of self replication. This is an example of emergent behavior from simple rules, and another example of the sentence "More is different". But this also raises the question: what is life? How to define it?

In the rest of this introduction, we will have the following approach: the game of life displays emergent behavior with simple rules. Since life itself appears as an emergent phenomenon, the aim of a chemist or physicist is to find out the underlying rules and how they appear. This is the approach we will have in this thesis.

The question of the definition of life is of the utmost importance because, before speaking about origins of life, we must define it, as said in ref [32], "Without a definition for life, the problem of how life began is not well posed". The common definition for a living organism implies three functions [33]:

- Replication: the ability to transmit information to offsprings.
- Metabolism: the ability to capture energy and transform it to stay away from thermodynamic equilibrium
- Compartmentalization: the ability to be a closed shell distinguishable from the environment

To widen this definition and go beyond life on earth, physicists introduced an extended definition of life called Lyfe [34] based on four physical considerations.

Although life is an emergent behavior due to increasing complexity, one of the approaches to understand its origins is to start from simple molecules and hypothesis and try to follow up until the basic chemical mechanisms of life are found in order to answer the question: how did these functions appear on earth? The first one to go in this way was Charles Darwin when he suggested that the evolutionary process started in a "warm little pond" [35]. This was then followed by A. I. Oparin and J. B. S. Haldane [36, 37], who independently proposed a scenario for the chemical evolution from small constituent of the atmosphere to biomolecules. Among these biomolecules are proteins and DNA/RNA molecules. The first ones have many functions in a cell, such as catalyst of reactions necessary for the metabolism of cells, or a structural role to hold the cell together [38]. Therefore, they fulfill the last two functions of the definition of life. The Replication part is carried out by the DNA/RNA molecules that hold the genetic code and hence the information to replicate a cell.

In the following of this thesis, we will be more interested in the proteins and more particularly to the building blocks of proteins. We will go once again down the scale ladder to simpler molecules: proteins are chains of small molecules called "amino-acids". There are 22 of them, and in this thesis we will be interested in the simplest of them: glycine.

The research around prebiotic formation of glycine is very active. One of the key works in this field is the well known Miller-Urey experiment [1], in which S. Miller gathered what was thought to be the atmosphere at the time of prebiotic earth (water (H_2O), methane (CH_4), ammonia (NH_3) and hydrogen (H_2)) and put a spark in it. He observed among other products the formation of glycine. The proposed mechanism for its formation is the Strecker one that was thought a century before [39]. Since this experiment, others have been performed using sparks and obtained the formation of more complex molecules [40, 41, 42, 43].

However, the Miller atmosphere is not the only way to obtain prebiotic molecules; there are many ways in which bio-molecules could have appeared on earth, and it is likely that it is a mix of all the scenarios that really happened. One of the scenarios is that life building blocks were brought to earth from the interstellar medium: they were either formed in comets [45, 46] or in interstellar ice via UV radiation [47, 48]. Other works show that elementary biomolecules were formed on earth in a primordial soup, hydrothermal submarine vents [49] or mineral surfaces [50, 51, 52]. A summary of all the possible environments is represented in figure 4 and taken from [44].

Nonetheless, due to the high number of possibilities, it is impossible to test all the possible conditions of formation of different molecules in the lab, as Miller did for a small set of molecules. To help experimentalists in their work, we can go down the scale ladder and go to the quantum scale. Using quantum mechanics, we can study chemical bond

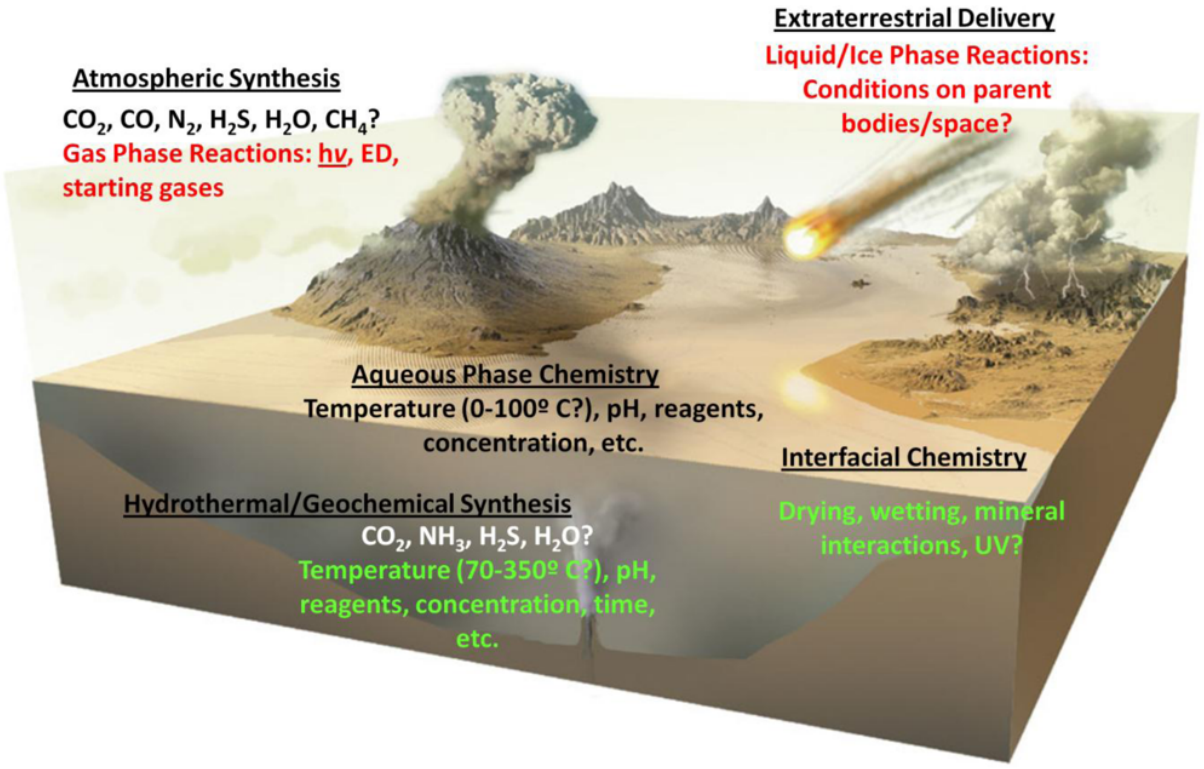

Figure 4: Artist representation of primordial earth and the different possible scenarios, adapted from [44]

formation and destruction under different conditions.

In his program, P. Dirac stated in 1929 [53] that the major laws of quantum mechanics were known but were too complex to be soluble for many atoms and that it was therefore "desirable that approximate methods of applying quantum mechanics should be developed". A solution to this question came with the computer and the approximate resolution of the Schrödinger equation. *Ab initio* calculations are a solution to that question because they are an approximation of the solution of the Schrödinger equation for many electrons that came in the 1960s. They are called *ab initio* because they only depend on the Schrödinger equation and do not need any external empirical parameter. With such computations, it is possible to simulate the dynamics of a system and take into account the electronic degrees of freedom, and thus study the thermodynamics and kinetics of a chemical reaction.

Indeed, the aim of such studies is to first understand the transition mechanism of the reaction and identify all the intermediates: some of them are too short-lived to be observed in experiments and are only hypothesized. Simulations allow confirming these hypotheses. The other goal of *ab initio* studies is to get thermodynamical and kinetics information. Often, the free energy difference and the activation barrier are obtained. The free energy difference tells which of the reactants or the products is the most stable, while the activation barrier gives information on the feasibility of the reaction, since it is

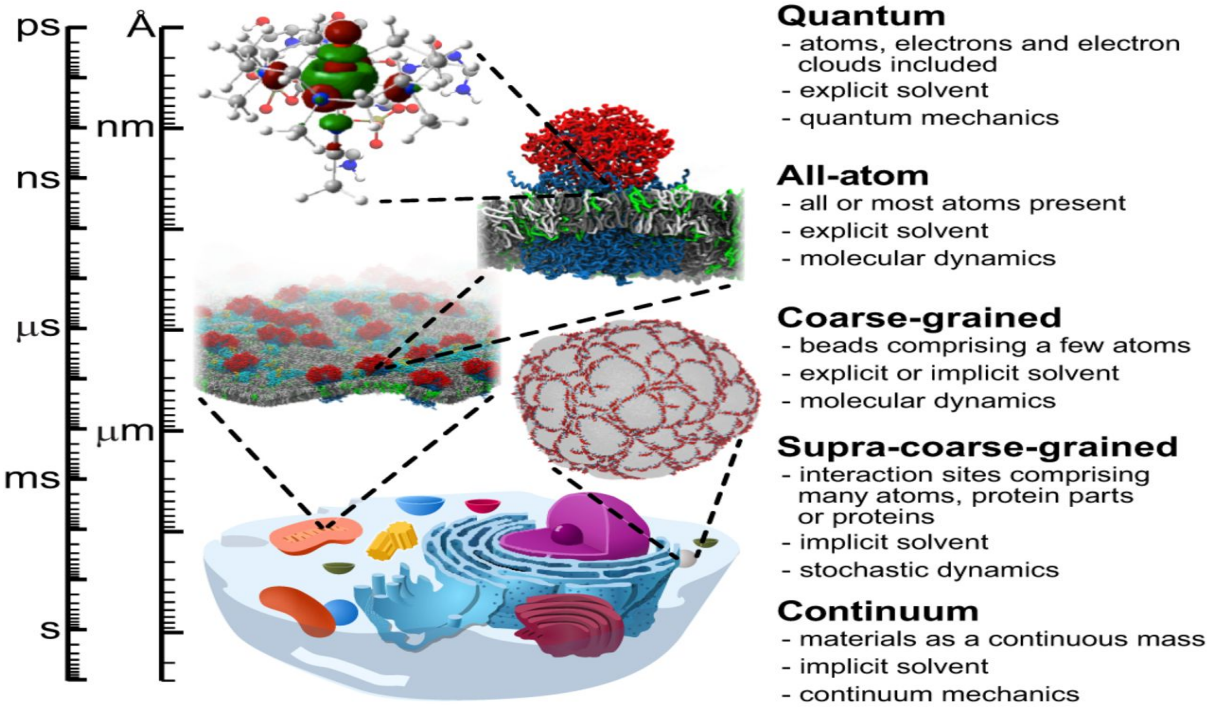

Figure 5: Illustration of different scales and the simulation techniques used, adapted from [62]

linked to the kinetic rate. They can be compared to the available experimental data via the equilibrium constant and the measured kinetic rate.

For example, one of the key studies in this field is the simulation of the Miller experiment in a simulation box [54], where the products observed by Miller were also observed in the simulation, and the transition mechanisms were accessible contrary to the experiment. Some very accurate simulations are also performed in gas phase to see the behavior in the interstellar medium and the possible formation of biomolecule from ices and dust [55, 56, 57]. Some other studies perform calculations in solution to see the formation of molecules in the "warm little pond" of Charles Darwin [6, 58, 59, 60] and understand the mechanisms of formation as well as the kinetic and thermodynamics.

In the team, a protocol has been set in place to overcome the intrinsic bottleneck of the transition time scales that are not reachable by any means of simulation. To do so, enhanced sampling techniques such as metadynamics [2] and Umbrella Sampling [4] are used along with state of the art collective variables [61]. This led to thorough studies of prebiotic scenarios on early earth, and to the whole step by step study of the Strecker synthesis of glycine [7].

However, *ab initio* molecular dynamics simulations have limitations: they are slow and scale with the cube of the number of electrons which is a limit for the time-scales and length-scales which can be studied with *ab initio* studies, as shown in figure 5. This makes the quantitative sampling phase of the protocol of the team very expensive in terms of

computational time, and therefore, all the possible conditions of the prebiotic earth cannot be studied, for example, it is difficult to include mineral surfaces, water and reactive molecules in a simulation box. It is also hard to study large reactive systems such as for example polymerized glycine in water using *ab initio* simulations

To go further in Dirac's program and to remove the bottleneck of computational time, one solution can be to use: artificial intelligence or machine learning, as it has allowed great progress in other fields such as protein folding [63]. In molecular dynamics simulations, several methods have been introduced [9, 8, 64, 65] to reduce the computational time. As the expensive part is to compute the energies and forces, a model is trained with an initial dataset containing the positions of the atoms of the system and the associated energies and forces.

These methods have been used to study equilibrium systems [11, 12, 13], to expand the number of atoms in a simulation box while keeping the *ab initio* accuracy. But the question of using machine learning methods for reactive systems was an open question at the beginning of my thesis, although some progress has been made by other teams [14, 15].

The main problem of machine learning models for molecular dynamics simulations is the extrapolation problem: how can one trust the prediction of the model, and how can one keep it away from configurations it does not know. This question is partly related to the question of choosing training points: for a model of a reactive system, we need a model that has a correct behavior all along the chemical space, and therefore, training point must be carefully chosen. Indeed, it is useless to train a model if all the data needed from the model is already in the training set. The right balance should thus be chosen between a sparse training set with not enough energies and a huge training set with redundancy in information and expensive to generate.

The aim of this thesis is to provide methods to build a training set and applying it to different chemical reactions in solution.

In this manuscript, we will start by presenting the methods used to study chemical reactions in solutions in the team, with an application to the Strecker synthesis of glycine. We will then explain in the second chapter the basic principles of machine learning and how it can be applied to molecular dynamics simulations. In a second part, we show the first method devised in this thesis to build a training set for a model for chemical reactions in solution. We first build this method by using the data of the first step of the Strecker synthesis of glycine and obtain results with an accuracy comparable to the *ab initio* one. We apply this method to the third step of the same mechanism, which is more complex and still obtain satisfactory results. We thus apply this method to a new pathway towards glycine from the same precursors as in the Strecker one. This pathway has never been studied, and we manage to explain some experimental results obtained. Finally, in the last part of this thesis, we point out that the previous method depends on the prior knowledge of the transition mechanism. We thus present a last methodological part introducing tools to explore transitions without any prior knowledge of the mechanism and apply it for the

training of machine learning potential for chemical reactions in solution. In this chapter, we use a very simple benchmark reaction that is the S_N2 substitution of methyl-chlorine with a chlorine ion [66].

Part I

Methods: Study of chemical reactions in solution and basic principles of machine learning potentials

Chapter 1

Ab initio study of chemical reactions in solution

1.1 Quantum mechanics

In order to study chemical reactions in solution, electronic degrees of freedom need to be treated to account for the creation or the destruction of chemical bonds. To do so a variety of methods have been devised, they are called *ab initio* because they are derived from first principle and don't need external empirical parameters. Among these methods, density functional theory (DFT) allows a good tradeoff between accuracy and affordable computational time for the systems we are interested in. In the following section, we will first present the historical Hartree-Fock (HF) approach [67, 68] that will allow us to introduce many concepts that will be used throughout this thesis, such as self-consistent resolution of equations and variational principles. The latter has also a big importance in machine learning, this is why, even though the HF method was not used in this thesis we chose to present the basic principles of this method.

1.1.1 The quantum many body problem

In the study of the quantum behavior of a system with N atoms of atomic number $(Z_i)_{i=1..N}$ with N_e electrons with positions $\mathbf{r_1}, ..., \mathbf{r_{N_e}}$ and N_p nuclei with positions $\mathbf{R_1}, ..., \mathbf{R_{N_p}}$, one wants to find the global wavefunction, $\Psi(\mathbf{r_1}, ..., \mathbf{r_{N_e}}, \mathbf{R_1}, ..., \mathbf{R_{N_p}})$ associated to the ground state of a system with its corresponding energy. Because all the relevant information can be extracted from the wavefunction, energies and forces. This is performed by solving the time dependent Schrödinger equation:

$$i\hbar \frac{\partial \Psi(\mathbf{r_1}, ..., \mathbf{r_{N_e}}, \mathbf{R_1}, ..., \mathbf{R_{N_p}})}{\partial t} = H\Psi(\mathbf{r_1}, ..., \mathbf{r_{N_e}}, \mathbf{R_1}, ..., \mathbf{R_{N_p}}) \tag{1.1}$$

Where H is the Hamiltonian of the system that can be separated in the following form:

$$H = T_N + V_{N,N} + V_{N,e} + T_e + V_{e,e} \tag{1.2}$$

with T_N the total kinetic energy associated to the nuclei given by, with M_i, the mass of nucleus i

$$T_N = -\frac{\hbar^2}{2}\sum_{i=1}^{N_p}\frac{\nabla_{R_i}^2}{M_i} \tag{1.3}$$

$V_{N,N}$ the potential energy associated to the coulombic interaction between nuclei:

$$V_{N,N} = \sum_{i=1}^{N_p}\sum_{\substack{j=1\\ j>i}}^{N_p}\frac{Z_i Z_j e^2}{|\mathbf{R_i}-\mathbf{R_j}|} \tag{1.4}$$

$V_{N,e}$ the term corresponding to the coulombic interaction between nuclei and electrons:

$$V_{N,e} = -\sum_{i=1}^{N_p}\sum_{j=1}^{N_e}\frac{Z_i e^2}{|\mathbf{R_i}-\mathbf{r_j}|} \tag{1.5}$$

T_e is the total kinetic energy of the electrons:

$$T_e = -\frac{\hbar^2}{2m_e}\sum_{i=1}^{N_e}\nabla_{r_i}^2 \tag{1.6}$$

and finally $V_{e,e}$ is the coulombic interactions between electrons given by:

$$V_{e,e} = \sum_{i=1}^{N_e}\sum_{\substack{j=1\\ j>i}}^{N_e}\frac{e^2}{|\mathbf{r_i}-\mathbf{r_j}|} \tag{1.7}$$

The time dependent Schrödinger equation can be simplified into an eigenvalue problem by assuming stationary states:

$$H\Psi(\mathbf{r_1},...,\mathbf{r_{N_e}},\mathbf{R_1},...,\mathbf{R_{N_p}}) = E\Psi(\mathbf{r_1},...,\mathbf{r_{N_e}},\mathbf{R_1},...,\mathbf{R_{N_p}}) \tag{1.8}$$

This eigenvalue problem is nonetheless impossible to solve for realistic system and many approximations based on the physics of the system have been introduced throughout the history of quantum mechanics. The first approximation that we will use here is the Born-Oppenheimer approximation [69]. It was first used to solve the hydrogen atom problem. The idea behind this approximation is that the electrons are much lighter than the nuclei, in other words, from the point of view of the electrons, nuclei are not moving and can therefore be considered as classical particles allowing a decoupling between the two parts. All the terms in the Hamiltonian regarding purely nuclear interactions can thus be put into a constant that we will write from now on $E_{N,N}$. The Hamiltonian of the system then becomes:

$$H = E_{N,N} + T_e + V_{e,e} + V_{e,N} \tag{1.9}$$

Where the term $E_{N,N}$ can be discarded as it is a constant classical term. This problem is still a many-body problem where electrons interact with each other. It can only be solved analytically for the hydrogen atom and hydrogen-like atoms called hydrogenoids

because the number of dimensions of the problem grows exponentially with the number of electrons. This is also what Dirac intended in his program cited in the introduction of this thesis [53]. Nonetheless, in a simulation box involving hundreds of electrons, the solution of this equation cannot be found, and one has to introduce physically inspired approximations to obtain accurate solutions of the Schrödinger equation.

In the following sections, we will first introduce the historical mean-field Hartree-Fock approximation, which is a common way of solving many-body problems that helps circumvent the problem of the growing number of dimensions in the many body problems. In these approximations, the electrons behave as non-interacting particles that evolve in a potential created by the other neighboring electrons.

1.1.2 Hartree-Fock formalism

In the HF approximation, we look for a separable wavefunction, *i.e.*, a wavefunction that can be written as the product of single-particle wavefunctions, therefore we introduce a trial wavefunction

$$\Phi(r_1, r_2, ..., r_{N_e}) = \phi_1(r_1)\phi_2(r_2)...\phi_{N_e}(r_{N_e}) \tag{1.10}$$

Which leads for particle i to the following eigenvalue problem:

$$h(\mathbf{r_i})\phi_i(\mathbf{r_i}) = \epsilon_i\phi_i(\mathbf{r_i}) \tag{1.11}$$

With $h(\mathbf{r_i})$ given after some algebraic manipulations by:

$$h(\mathbf{r_i}) = -\frac{\hbar^2}{2m_e}\nabla^2 - \sum_{j=1}^{N_p}\frac{Z_j e^2}{|\mathbf{r_i} - \mathbf{R_j}|} + e^2\sum_{k\neq i}^{N_e}\int\frac{|\phi_k(\mathbf{r}')|^2 d\mathbf{r}'}{|\mathbf{r_i} - \mathbf{r_k}|} \tag{1.12}$$

Even though equation 1.15 is a single body equation, the other particles play a role in the "bath" term of h in equation 1.12. This means that to know $\phi(\mathbf{r_i})$, the whole set of single-particle wavefunctions needs to be known. This is why this set of equation needs to be solved self-consistently: first we start by a guess of the solutions, we solve the set of equations 1.15 and update the set of single-particle wavefunctions. This is done until convergence on the energy is achieved, *i.e.*, the difference of ground state energy is below a defined threshold.

This scheme was introduced by Hartree [67], but, later on, Slater [70] and Fock pointed out that the solutions of 1.15 were not antisymmetric with respect to the exchange of two particles, which means that the solutions did not lead to a fermionic behavior. This is why they introduced a new ansatz for the trial wavefunction called a "Slater determinant":

$$\Phi(\mathbf{r_1}, ...\mathbf{r_{N_e}}) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \phi_1(\mathbf{r_1}) & \phi_2(\mathbf{r_1}) & \phi_3(\mathbf{r_1}) & \dots & \phi_{N_e}(\mathbf{r_1}) \\ \phi_1(\mathbf{r_2}) & \phi_2(\mathbf{r_2}) & \phi_3(\mathbf{r_2}) & \dots & \phi_{N_e}(\mathbf{r_2}) \\ & & \dots\dots\dots\dots\dots & & \\ \phi_1(\mathbf{r_{N_e}}) & \phi_2(\mathbf{r_{N_e}}) & \phi_3(\mathbf{r_{N_e}}) & \dots & \phi_{N_e}(\mathbf{r_{N_e}}) \end{vmatrix} \tag{1.13}$$

where, the functions ϕ_i are no longer just functions of the spatial coordinates, but also the spin coordinate of electron i (namely a product of a spatial part and a spin part, and

since the Hamiltonian does not depend on the spin, the spin degrees of freedom can be integrated out).

In physics, problems are often solved by looking for a variational ansatz, this is also the case with this problem by introducing the following variational principle:

Variational principle: The energy of the ground state of a quantum system is determined by minimizing the energy with respect to the wave function, more formally, E_0, the energy of the ground-state of a system is given by:

$$E_0 = Min_\Psi \frac{\langle \Psi | H | \Psi \rangle}{\langle \Psi | \Psi \rangle} \tag{1.14}$$

In our case, we need to determine the Slater determinant that minimizes the ground state energy. To do so, Lagrange multipliers can be introduced, and the following eigenvalue equations are obtained:

$$h_{HF}(\mathbf{r_i})\phi_i(\mathbf{r_i}) = \epsilon_i \phi_i(\mathbf{r_i}) \tag{1.15}$$

the ϵ_i being the Lagrange multipliers, and h_{HF} is given by:

$$h_{HF}(\mathbf{r_i}) = -\frac{\hbar^2}{2m_e}\nabla^2 - \sum_{j=1}^{N_p} \frac{Z_j e^2}{r_{ji}} + V_{HF}(\mathbf{r_i}) \tag{1.16}$$

V_{HF} is the Hartree-Fock potential energy due to the interaction of electron i with the mean-field created by all the other electrons, it is given by the following equation:

$$V_{HF}(\mathbf{r_i}) = \sum_{k=1}^{N_e} (J_k(\mathbf{r_i}) - K_k(\mathbf{r_i})) \tag{1.17}$$

The term $J_k(\mathbf{r_i})$ is the Coulomb operator which corresponds to the Coulomb repulsion felt by atom i and caused by all the other electrons and is given by:

$$J_k(\mathbf{r_i}) = e^2 \int \frac{|\phi_k(\mathbf{r'})|^2 d\mathbf{r'}}{|\mathbf{r_i} - \mathbf{r_k}|} \tag{1.18}$$

The other term is a purely quantum mechanical term created by the Pauli exclusion principle. It is called the exchange term and is given by:

$$K_k(\mathbf{r_i})\phi_i(\mathbf{r_i}) = e^2 \int \phi_k^*(\mathbf{r'}) \frac{1}{|\mathbf{r_i} - \mathbf{r'}|} \phi_i^*(\mathbf{r'})\phi_k^*(\mathbf{r_i}) d\mathbf{r'} \tag{1.19}$$

The eigenvalue equations are then solved self-consistently by decomposing each single-particle function on a chosen basis of functions, which then gives a system of equations for the coefficients of the basis functions.

We now have introduced a first method to characterize the ground state of a system using a variational principle and self-consistent resolution of non-linear equations. The variational principle showed that the ground state energy is a functional of the full Slater-determinant. In the next section, using the electron density function, we will show that it is possible to characterize the ground state of a system by seeing the ground state energy as a functional of the density function.

1.1.3 Density functional theory

In the previous section, we showed using a variational principle that the ground state energy of a system can be determined by minimizing the energy with respect to the wavefunction. This makes the ground state energy a functional of $4N_e$ variables if we take into account the spin, which makes it hard to manipulate. The complexity increases with the number of electrons. In the 1960s, Paul Hohenberg, Walter Kohn and Lu Sham introduced a way to express the ground state energy as a functional of the electronic density, which is a function of only 3 spatial coordinates. This method is called density functional theory, and Walter Kohn was awarded the chemistry Nobel Prize in 1998 for this method. In the following sections, we will introduce the basic principles of density functional theory. Since this thesis, involves many methodological tools, and DFT is a standard well known tool, we will not enter the details of the derivations of the equations but adopt a more practical view on the tools used during this thesis.

The density function

Instead of considering the probability distribution of each electron individually, DFT is based on the total electronic density. Since electrons are indistinguishable particles, we can evaluate the density probability of presence of an electron in a finite volume element $d\mathbf{r}$. We will refer to this quantity as the electronic density in the future:

$$n(r) = N_e \int |\Psi(\mathbf{r}, \mathbf{r_2}, \mathbf{r_3}, \dots \mathbf{r_{N_e}})|^2 d\mathbf{r_2} d\mathbf{r_3} \dots d\mathbf{r_{N_e}} \tag{1.20}$$

The key to density functional theory are the two Hohenberg-Kohn theorems [71]:

The Hohenberg-Kohn theorems

First Hohenberg Kohn theorem: For an interacting system of electrons, the external potential V_{ext} is fully and uniquely determined up to an additive constant by the electronic ground state density $n(\mathbf{r})$, the total ground state energy of the system is thus a functional of the electronic density.

According to this theorem, the ground state total energy of the system can thus be expressed as a functional of the electronic density:

$$E[n(\mathbf{r})] = T_e[n(\mathbf{r})] + V_{e,e}[n(\mathbf{r})] + V_{ext}[n(\mathbf{r})] \tag{1.21}$$

where V_{ext} is the interaction between the electrons and the nuclei, as nuclei are considered as classical external particles it can be written:

$$V_{ext}[n(\mathbf{r})] = -e^2 \int \sum_{j=1}^{N_p} \frac{Z_j}{|\mathbf{r} - \mathbf{R_j}|} n(\mathbf{r}) d\mathbf{r} \tag{1.22}$$

The same kind of expression can be obtained for the other terms, although this does not give us any indication on how to obtain the electronic density of the ground state. To do so, the second Hohenberg-Kohn theorem is used:

Second Hohenberg-Kohn theorem: For a given external potential V_{ext} the energy of the ground state is given by the global minimum of the energy functional $E[n(\mathbf{r})]$. Therefore, the knowledge of the functional $E[n(\mathbf{r})]$ is enough to determine the ground state and the electronic density of the system.

This is a second variational principle, that allows determining the properties of the system by minimizing the energy functional with respect to the electronic density.

The Kohn-Sham equations

Minimizing the energy functional is still a non-trivial task due to the multi-body terms in $V_{e,e}[n(\mathbf{r})]$. To overcome this problem, Walter Kohn and Lu Sham came up with an elegant solution [72] that was summed up by Richard M. Martin [73]: "If you don't like the answer, change the question". Since the Hohenberg-Kohn theorem states that a system is determined by its electronic density, the many-body problem can be mapped to a single-particle problem with the same electronic density, and therefore the ground state energy of this fictitious single-particle system will be the same as the one of the real system. The ground state energy functional of this fictitious system (denoted with subscript S) is:

$$E_S[n(\mathbf{r})] = T_S[n(\mathbf{r})] + E_H[n(\mathbf{r})] + E_{XC}[n(\mathbf{r})] + V_{ext}[n(\mathbf{r})] \tag{1.23}$$

With T_S the kinetic energy of S, E_H the Hartree functional accounting for coulombic interaction in a "mean-field" manner as seen in the previous section, it is the equivalent of the J_k term in equation 1.17. On the other hand, these two terms only account for single body potentials, this is where the term $E_{XC}[n(\mathbf{r})]$ comes into play and represents the deviation of the system from the single-body mean-field behavior.

$$E_{XC}[n(\mathbf{r})] = T[n(\mathbf{r})] + V_{e,e}[n(\mathbf{r})] - (T_s[n(\mathbf{r})] + E_H[n(\mathbf{r})]) \tag{1.24}$$

$E_{XC}[n(\mathbf{r})]$ must contain an exchange term, as the K_k term in equation 1.17 which comes from the Pauli exclusion principle as shown in the previous section, but also a correlation part that accounts for the interaction of particles of the same sign. It can then be shown that minimizing $E[n(\mathbf{r})]$ is equivalent to solving the following single particle Schrödinger equation:

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}} + V_{ext}(\mathbf{r}) + V_H(\mathbf{r}) + V_{XC}(\mathbf{r})\right]\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{1.25}$$

With ϕ_i the single-particle wavefunctions that are called orbitals. V_{ext} is the external potential and does not depend on the density, V_H is the monoelectronic coulombic interaction potential given by:

$$V_H(\mathbf{r}) = e^2 \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \tag{1.26}$$

And V_{XC} is the exchange correlation potential that we don't know exactly and that we have to approximate which is given by:

$$V_{XC}(\mathbf{r}) = \frac{\delta E_{XC}[n(\mathbf{r})]}{\delta n(\mathbf{r})} \tag{1.27}$$

The final density function is given by:

$$n(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2 \tag{1.28}$$

Therefore, if we put together equations 1.25 and 1.28 we have two equations that we need to solve self-consistently: first start by a guess of the electronic density, then solve equation 1.25, then compute the new density, and iterate until convergence is achieved on the energy, *i.e.*, when the difference in the ground state energy between two steps of the procedure is below a pre-defined threshold.

The exchange-correlation functional:

The exchange-correlation functional, $E_{xc}\left[n(\mathbf{r})\right]$ is the key approximation of DFT. We have no way of knowing it analytically. This is why there are several ways to approximate it, and the level of approximation will impact the accuracy of the predictions. However, it also increases the computational time, therefore a tradeoff between accuracy and feasibility needs to be found as a function of the system we want to study.

The local density approximation (LDA): The most simple approximation to make is to hypothesize that the exchange correlation functional can be approximated as an integral over the whole space of the exchange correlation energy per electron of a homogeneous electron gas ϵ_{xc}^{hom}:

$$E_{XC}[n(\mathbf{r})] = \int n(\mathbf{r})\epsilon_{xc}^{hom}(n(\mathbf{r}))d\mathbf{r} \tag{1.29}$$

This ϵ_{xc}^{hom} is computed using very accurate quantum Monte-Carlo simulations [74]. This approximation is one of the first ones introduced, it works well to compute properties of solids with electronic densities similar to the one of a homogeneous electron gas, *i.e.*, systems where the density varies slowly in space, crystals especially if they are metallic for example [75, 76]. It has however trouble reproducing properties of isolated molecules or systems with strongly correlated electrons. It is also known to overestimate the hydrogen bonding [77], which for us is crucial since we want to study chemical reactions in water where hydrogen bonding play an important role to stabilize transition states for example.

Generalized gradient approximation (GGA): A more refined way of approximating this function is by assuming it is not only a function of the local density, but also of the gradient, this way, the fluctuations of the electronic density are taken into account. In this case the exchange correlation functional is written:

$$E_{XC}^{GGA}[n(\mathbf{r})] = \int n(\mathbf{r})\epsilon_{xc}^{GGA}(n(\mathbf{r}), \nabla_{\mathbf{r}} n(r\mathbf{r}))d\mathbf{r} \tag{1.30}$$

There exists many GGA functionals, in this work we only used the Perdew-Burke-Ernzerhof (PBE) [78] functional that has the advantage of being completely *ab initio*, *i.e.*, it does not depend on any external parameter. It is worth mentionning that some GGA functionnals integrate the full HF exchange expression (K_k in equation 1.17) to increase accuracy, at the cost of increased computational cost. This family of functional are called hybrid functionals, such as B3LYP [79, 77].

The plane wave expansion

Now that we have a way to approximate the exchange correlation functional, we can insert the functional we want in equation 1.25 and solve it by expanding the single-particle wavefunction on a given basis, as it is done in the Hartree-Fock method. In the Hartree-Fock method, Gaussian basis functions are often preferred while in DFT, the chosen basis set is the plane wave basis. The reason for this is that DFT is mainly used to study periodic solids [80, 81] or liquids with periodic boundary conditions [82]. Therefore, the Bloch theorem can be applied: the single-particle wavefunctions can be written:

$$\phi_i(\mathbf{r}) = u_{i,\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}.\mathbf{r}} \tag{1.31}$$

and $u_{i,\mathbf{k}}$ has the same periodicity as the crystal lattice or the simulation box for a liquid for every $\mathbf{k}$ point in the first Brillouin zone. Since these functions are periodic, we can make a Fourier transform and the single-particle wavefunctions can be written:

$$\phi_i(\mathbf{r}) = \sum_{\mathbf{G}} c_{i,\mathbf{k},\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G}).\mathbf{r}} \tag{1.32}$$

With $\mathbf{G}$ a wave-vector of the Fourier transform. Finding all the $c_{i,\mathbf{k},\mathbf{G}}$ is however impossible. To a vector $\mathbf{G}$ one can associate a kinetic energy: $E = |\mathbf{G}|^2\hbar^2/2m_e$, and we can therefore compute the expansion for energies up to a certain energy cutoff that should be carefully chosen.

Pseudopotentials

Chemically speaking, the core electrons do not play a big role in the reactivity and are pretty static, taking them into account would make the computational cost grow. Moreover, close to the core, the valence electrons wavefunctions have big oscillations due to the Pauli exclusion principle, thus, a large amount of plane waves need to be used to describe these oscillations. To tackle this problem, pseudopotentials have been introduced. Plane waves are only used to model valence electrons, core electrons and the ionic potential are replaced by a pseudo potential that acts on the valence wavefunctions and reproduces the all-electrons wavefunctions after a certain cutoff. In this work, we only used the Martins-Troullier pseudopotentials [83].

Practical considerations when doing DFT calculations

Now that we have explained all the concepts of DFT, one has to take into account practical considerations, indeed, when performing atomistic simulations there is always a tradeoff to find between accuracy and computational time. Indeed, LDA calculations are cheap to run but not accurate for every system, on the opposite side, hybrid calculations between GGA and Hartree-Fock are very accurate but very expensive. Therefore, the exchange and correlation functional should be chosen as a function of the needs and the means, this tradeoff was compared with the Jacob ladder by John Perdew [84]. Moreover, even with the cheapest exchange correlation functional, big systems with more than a thousand atoms cannot be studied. This is why machine learning methods have been introduced to overcome this, and it is the subject of the third chapter.

1.2 Molecular dynamics

1.2.1 The ergodic hypothesis and probability distributions

First, statistical mechanics relies on a strong historical hypothesis: the ergodic hypothesis. Let us say we want to know the average of observable O. If we observe the evolution of the system for "a long enough time", making a time average is the same as performing an ensemble averaging. More formally, this can be written:

$$\lim_{\tau\to\infty}\int_0^\tau O(t)dt = \int p(\mathbf{x},\mathbf{p})O(\mathbf{x},\mathbf{p})d\mathbf{x}d\mathbf{p} \tag{1.33}$$

This hypothesis lead to the rise of statistical mechanics, because from this hypothesis, a statistical approach can be used to explain experimental results. In the rest of this manuscript, unless stated otherwise, we will be working in the canonical ensemble (N,V,T) with the canonical distribution for microstate $\mathbf{x}$ at inverse temperature $\beta = 1/k_BT$:

$$p(\mathbf{x}) = \frac{1}{Z}e^{-\beta E(\mathbf{x})} \tag{1.34}$$

With Z the partition function:

$$Z = \int e^{-\beta E(\mathbf{x})}d\mathbf{x} \tag{1.35}$$

When studying a transformation between state A and B, one often wants to know which of the state is the most stable, this is done by comparing the probabilities of being in one of the states:

$$p_{A/B} = \int_{\Omega_{A/B}} p(\mathbf{x})d\mathbf{x} \tag{1.36}$$

With $\Omega_{A,B}$ the region where states A and B are defined. When having a time evolution of the system, thanks to the ergodic hypothesis, $p(\mathbf{x})d\mathbf{x}$ is computed by performing a time average:

$$p(\mathbf{x}) = \frac{1}{\tau}\int_0^t au\delta(\mathbf{x}-\mathbf{x}(t))dt \tag{1.37}$$

To obtain this time evolution, molecular dynamics (MD) has been widely used to study phenomena at various scales from the quantum scale with material properties ($\mathring{A}$/nm) to macro proteins (μm)). In this section we explain the basic principles of MD and how to sample a system with a constant number of particles, volume and temperature: the NVT ensemble. This is the ensemble mainly used in this work. In this part, we will consider a system with N particles of mass $(m_i)_{i=1...N}$ and positions $(\mathbf{r_i})_{i=1...N}$ our goal is to obtain the time behavior of the system given the sum of forces $(\mathbf{F_i})_{i=1...N}$ acting on each particle of the system.

1.2.2 The Verlet algorithm

The most straight forward way to obtain the time evolution is by solving the Newton's equations of motions:

$$m_i\ddot{\mathbf{r}}_\mathbf{i} = \mathbf{F_i} \tag{1.38}$$

Where the double dot over $\mathbf{r_i}$ indicates the double time derivative of $\mathbf{r_i}$. This equation is however a continuous equation of time that cannot be solved analytically, moreover we only can propagate discrete equations.

Our goal is to obtain the time evolution of the system. From now on, we will call it the "trajectory" from $t = 0$ to $t = \tau$. To do so, we split this time interval into a discrete set of time lags: $(t_0, t_1, t_2 \ldots, t_{M-1}, t_M)$, with $t_0 = 0$ and $t_M = \tau$ and $t_{i+1} - t_i = \delta t$ with δt called the time step. Now, since we have the time evolution of $\ddot{\mathbf{r}}_\mathbf{i}$, and we want the time evolution of $\mathbf{r_i}$, we can write a Taylor expansion of $\mathbf{r_i}(t + \delta t)$ and $\mathbf{r_i}(t - \delta t)$ at the second order in δt that makes $\ddot{\mathbf{r}}_\mathbf{i}$ appear:

$$\mathbf{r_i}(t + \delta t) = \mathbf{r_i}(t) + \delta t \dot{\mathbf{r}}_\mathbf{i}(t) + \frac{\delta t^2}{2}\ddot{\mathbf{r}}_\mathbf{i}(t) + \frac{\delta t^3}{6}\dddot{\mathbf{r}}_\mathbf{i}(t) + O(\delta t^4) \tag{1.39}$$

$$\mathbf{r_i}(t - \delta t) = \mathbf{r_i}(t) - \delta t \dot{\mathbf{r}}_\mathbf{i}(t) + \frac{\delta t^2}{2}\ddot{\mathbf{r}}_\mathbf{i}(t) - \frac{\delta t^3}{6}\dddot{\mathbf{r}}_\mathbf{i}(t) + O(\delta t^4) \tag{1.40}$$

Thes two equations can then be summed to obtain:

$$\mathbf{r_i}(t + \delta t) = 2\mathbf{r_i}(t) - \mathbf{r_i}(t - \delta t) + \frac{\delta t^2}{2}\ddot{\mathbf{r}}_\mathbf{i}(t) + O(\delta t^4) \tag{1.41}$$

This equation does not depend on the velocities, and they can be obtained by subtracting equations 1.40 and 1.39. A more advanced scheme like velocity Verlet can be used to obtain the velocities. This method was introduced by Loup Verlet to study a Lennard-Jones system [85]. It does only depend on one parameter that is the time-step δt, and the error made is of the order of δt^4. Therefore, this parameter should be chosen with the system, it should not be too large to avoid missing the fast degrees of freedom of the system, but, if it is too small, the computational time needed to study a physical phenomenon will be too high. Hence, when performing molecular dynamics simulations, a tradeoff between a small and a bigger time-step must be found.

We thus have a way to obtain the positions of the particles of the system as a function of time, to then compute averages of thermodynamical quantities.

1.2.3 The Nosé-Hoover thermostat

To compute averages of thermodynamical quantities, one must choose the statistical ensemble in which this average is computed. As the Verlet algorithm conserves the total energy of the system, the natural ensemble of molecular dynamics is the (N, V, E) ensemble, where the number of particles, the volume of the system and the total energy of the system are conserved. But, due to for example high forces the temperature might have big variations, and one might prefer to work in the (N, V, T) ensemble by using a thermostat.

First, the temperature in statistical physics is defined at time t by:

$$\frac{3}{2}Nk_BT = \sum_{i=1}^{N}\frac{1}{2}m_i\dot{\mathbf{r}}_\mathbf{i}(t)^2 \tag{1.42}$$

Therefore, temperature is intrinsically related to the velocities of the system, hence if one wants to control the temperature, one has to control the velocities of the particles in the system. To do so, a thermostat is used. There exists many thermostats, in this thesis, we only used the Nosé-Hoover thermostat [86, 87] that we will present now.

In this thermostating method, an additional degree of freedom χ is introduced that accounts for the exchange of energy between the thermostat and the particles of the system, the equations of motion become:

$$m_i \ddot{\mathbf{r}}_\mathbf{i}(t) = \mathbf{F_i} - m_i \chi \dot{\mathbf{r}}_\mathbf{i}(t) \tag{1.43}$$

With the time evolution of χ given by:

$$\dot{\chi} = \frac{1}{Q}\left[\sum_{i=1}^{N} m_i \dot{\mathbf{r}}_\mathbf{i}^{\,2} - 3Nk_BT_0\right] \tag{1.44}$$

This additional degree of freedom can be thought of as a friction term in the system: if $T > T_0$, then, the friction term increases and takes energy from the system, while if $T < T_0$, then, the friction term gives back energy to the system in order for the instantaneous temperature to oscillate around the target temperature T_0.

1.2.4 Periodic boundary conditions

In order to avoid boundary effects in the simulation box, periodic boundary conditions are used. This means the system is periodically repeated in the three directions of the system. More specifically, when an atom leaves the simulation box, it enters back the box by the other side. In the case of a cubic simulation box of size a, the coordinate α of atom i becomes when it leaves the box:

$$r_i^\alpha(t) = r_i^\alpha - floor(r_i^\alpha/a) \tag{1.45}$$

By using periodic boundary conditions, first we exactly control the volume of the simulation box, and second, we can use the plane wave expansion of DFT explained in the previous section.

1.2.5 Obtaining the forces

Now that we have equations to get the motion of the atoms, we need to get a way to have the forces acting on each atom. This, is the last piece to have a working MD simulation.

Classical molecular dynamics

The easiest way to perform MD simulations is with so-called classical MD simulations, where the forces are derived from parameterized force fields. These parameters are either fit from experimental data, or fit from more accurate *ab initio* calculations. They are less accurate than *ab initio* simulations but allow studying bigger time and length scales, since they often linearly scale with the number of atoms. For example, for aqueous systems, the most used force fields are TIP3P/TIP4P. [88, 89] The historical first force field that

will be used in the next section to illustrate the concept of free energy barrier is the Lennard-Jones potential:

$$E_{pot} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.46}$$

Where ϵ_{ij} models the strength of the interaction between atom i and atom j and σ_{ij} the equilibrium distance between atom i and atom j. It is often used to model the behavior of noble gas.

ab initio molecular dynamics

With classical MD, however, all the information about the electronic behavior of the system is lost. Therefore, no chemical reaction can happen during a classical molecular simulation. This is why *ab initio* molecular dynamics (AIMD) simulations are performed for chemical reactions. We will present here how it is done in the case of the Born-Oppenheimer approximation: nuclei are treated as classical point particles on which electronic forces act. The force acting on the nucleus at the position $\mathbf{R_i}$ is given by the Hellman Feynman theorem:

$$\mathbf{F_i} = \langle \Psi | \nabla_{\mathbf{R_i}} H_e | \Psi \rangle - \nabla_{\mathbf{R_i}} E_{N,N} \tag{1.47}$$

This is with this last equation that it can be seen that performing an AIMD simulation is very expensive in computational time, because the DFT equations need to be solved at every time step. We now have introduced every tool to perform simulations to gather thermodynamical information about a system we want to study. In the next section, we will see how to sample the phase space in order to get the relevant information about the system.

1.3 Computing free energies

In this section, we will make the link between statistical physics, ensemble averages, molecular dynamics and experimental measurements. To do so, we will use a toy model that is a Lennard-Jones dimer immersed into a Lennard-Jones solvent. The interaction energy is bigger for the particles of the dimer than for the solvent particle but also for the interaction between the solvent particles and the dimer particles. We wish to study the association/dissociation process of the dimer particles. With this toy model, we will illustrate all the relevant concepts of sampling events.

1.3.1 Free energy and collective variables

The average in equation 1.37 can be performed for very small systems, because the volume of the phase space hypersphere of dimension 3N grows exponentially with N. This means that it is impossible to compute for realistic systems. To compute this probability, we need to reduce the dimensionality of the system by projecting the dynamics of the system

onto the relevant degrees of freedom. This is done by using collective variables (CV).

A CV, or a reaction coordinate (RC) is a projection of the 3N dimensional atomic positions space onto a space of dimension n where n is smaller than 3N (often n is 1 or 2). For chemical reactions, a good CV should be a quantity that allows us to identify the start and end states, but also that captures the transition mechanisms. It is this last point that is the hardest to achieve, and that makes the difference between a CV and an order parameter. Let us name the collective variable $s(\mathbf{x})$, we define Ω_A as $s \in [s_A^0, s_A^1]$ and Ω_B as $s \in [s_B^0, s_B^1]$, the reactants and products spaces. In this framework, we can introduce the marginalized probability density:

$$p(s) = \int P(\mathbf{x})\delta(s - s(\mathbf{x}))d\mathbf{x} \tag{1.48}$$

and equation 1.36 becomes:

$$p_{A,B} = \int_{s_{A/B}^0}^{s_{A/B}^1} p(s)ds \tag{1.49}$$

We also define the marginalized free energy, also called free energy profile or surface:

$$F(s) = -\frac{1}{\beta}\log p(s) \tag{1.50}$$

1.3.2 Thermodynamics from free energy profiles:

The last step is to make the link between our microscopic calculations and the macroscopic quantities experimentally measured. Using the free energy profiles, important thermodynamic quantities may be defined to compare with experiments. First, in experiments, the equilibrium constant of a reaction, K, is often measured. For a liquid phase reactions such as:

$$\mathrm{aA} + \mathrm{bB} \rightleftharpoons \mathrm{cC} + \mathrm{dD} \tag{1.51}$$

The equilibrium constant is defined as:

$$K = \frac{[C]^c[D]^d}{[A]^a[B]^b} \tag{1.52}$$

where the square brackets indicate the concentration of the species. From this definition, it is natural that if $K < 1$, the reaction is not thermodynamically favored, while if $K > 1$ it is. These macroscopic thermodynamics considerations can be linked to our microscopic free energy calculations via the following relation:

$$\Delta F^0 = -k_B T \ln K \tag{1.53}$$

This allows to make the link between simulations and experiments and compute the relative stability of the reactants and the products. It is however more difficult to compute the kinetic rates using molecular dynamics simulations, and it will be discussed in detail in chapter 5. One simple qualitative way to compare the kinetics of competitive mechanisms is to compute the activation free energy. Indeed, according to transition state theory

(TST) [90], the kinetic rate is proportional to the exponential of the opposite of the activation energy. In other words, the higher the activation free energy, the less kinetically favorable the transition. As every mechanism is a competition between kinetics and thermodynamics, it is important to discuss both aspects.

1.3.3 Application to a Lennard-Jones system

Now that all the theoretical framework has been put into place, we can illustrate all the concepts for a very simple system, before going into more complex phenomena such as AIMD study of chemical reactions. To do so, we take a Lennard-Jones dimer solvated in a Lennard-Jones solvent as mentioned in the introduction of this section.

We want to know the free energy difference between the associated state and the dissociated state. To do so, we used the molecular dynamics package LAMMPS that was also used in this thesis. For this simulation, the k_BT was set to the strength of the interaction between the two atoms of the dimer. The system is represented in figure 1.1, with the associated state where the atoms of the dimer are linked and the dissociated state where the atoms of the dimer are separated by at least one solvent atom.

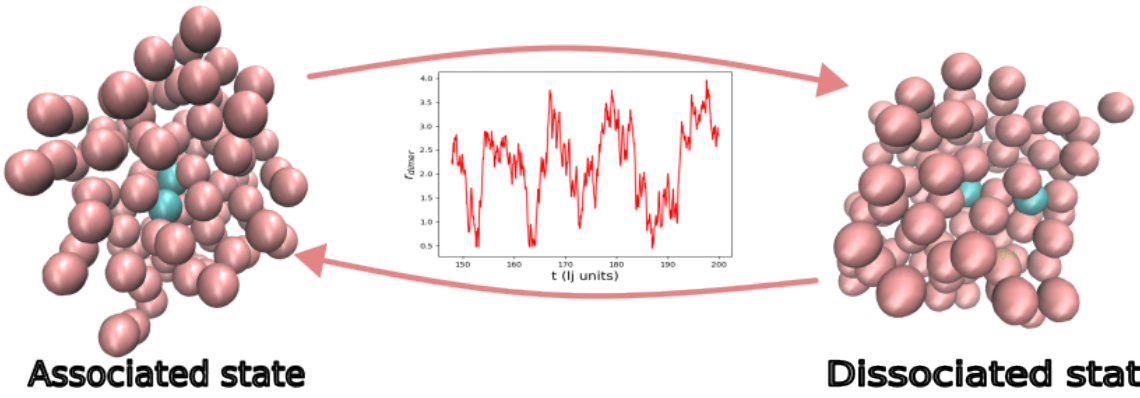

Figure 1.1: Toy model used in this section, the box contains 100 solvent atoms (pink) and two interacting atoms from the dimer represented in blue. We are looking at the time evolution of r_{dimer}, the distance between the two atoms of the dimer to characterize the associated and dissociated forms.

One can be interested in the mean pressure of the system or other physical observable, but since our final goal is to be able to study chemical reactions and the stability of a state, we will here be interested in the free energy. As indicated in figure 1.1 we will use as CV the distance between the two atoms of the dimer, get the probability histograms and build the free energy. To assess the accuracy of the free energy profile (FEP), the second half of the simulation is taken, cut into two parts and the probability histograms are computed on these two parts. If the free energies computed within these two parts are within k_BT we can safely say that convergence is achieved and keep the free energy. The mean of the two FEP will be the observable, while the difference between the two is

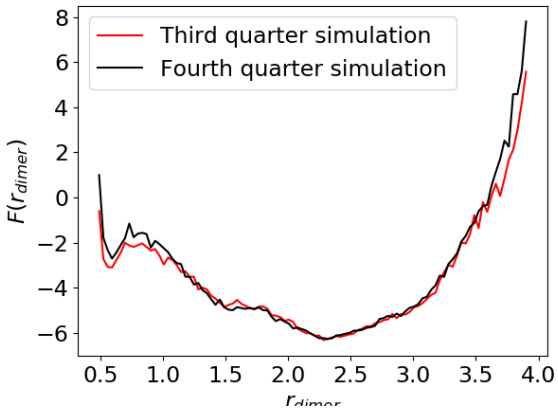

Figure 1.2: Free energy obtained for the dissociation process of a simple Lennard-Jones system, the two last quarters of the simulation are taken to compute to assess the uncertainty (in this case, around $0.5k_BT$)

the uncertainty.

The results can be seen in figure 1.2. We can draw several conclusions from this plot: first, three states can be identified: the bounded associated state, in the vicinity of $r_{dimer} = 0.5$, but this state is a very short-lived metastable state, since the barrier to leave this state is of the same order as the thermally induced fluctuations. Then, there is a plateau around $r_{dimer} = 1.5$ which corresponds to states where the dimer atoms are separated by one solvent atom, and finally, there is the completely free state after.

However, this is in the case where the interactions between the dimer atoms is comparable to the thermal energy, therefore the thermal agitation gives enough energy to cross the activation barrier. Let us see what happens when we increase the interaction energy but keep the temperature and the simulation time unchanged.

1.3.4 The case of rare events

In figure 1.3 we changed the strength of the interaction between the atoms of the dimer (ϵ_{dimer}) while keeping all the other parameters of the system unchanged (mainly the temperature and the simulation time). Qualitatively, we observe that the stronger the interaction is, the more stable the associated state is. Furthermore, as expected, the barrier between the associated state and the dissociated state also increases. From the convergence analysis that we can lead by looking at the last two quarters of the simulations, we conclude that the free energy of this dissociation process is well sampled until $\epsilon_{dimer} = 10$, then, the fluctuations in the simulation where $\epsilon_{dimer} = 20$ ($10k_BT$) between the two last quarters of the simulation are too high to conclude on the barrier height. This gets even worse when increasing ϵ_{dimer} to 40 Lennard-Jones units ($20k_BT$).

These results can be put in the framework of chemical reactions. Indeed, the simulation time is limited, and no ergodic simulation can be run for barriers higher than a few

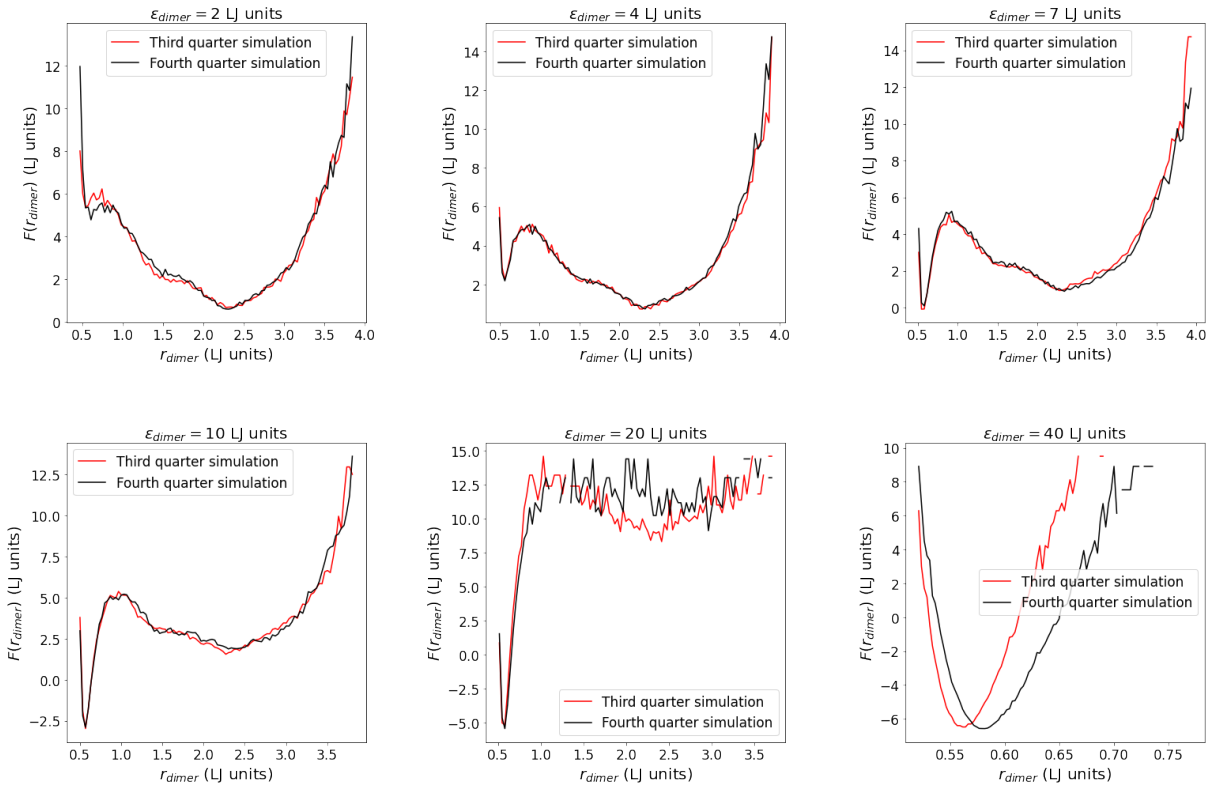

Figure 1.3: Evolution of the free energy profile of a simple Lennard-Jones (LJ) dimer in a solvent. The interaction strength of the dimer is changed from 2LJ units to 40 LJ units, while K_BT the thermal energy was set to 2LJ units. The simulation time was kept unchanged

k_BT (see figure 5 in the introduction). For example, proton exchanges have a barrier of around $20k_BT$ at room temperature, and covalent bond breakings have even higher barriers. Therefore, in this framework only the equilibrium basins can be studied, depending on the initial configuration, but there is almost no chance that the system crosses the barrier, and even so, we need many transitions to have correct probability histograms, as shown with the case of figure 1.3 where $\epsilon_{dimer} = 20$, where some transitions are observed but not enough to get correct convergence.

A good sampling of the transition state (TS) region, which is the region near the top of the barrier is needed to understand the whole complexity of the mechanism and its kinetics. This is however the most difficult task due to the short-lived nature of the TS: if a simulation is started around the TS it will either end up, in the reactants basin or in the products basin. We will get back to this definition in chapter 5. To overcome this simulation time bottleneck, enhanced sampling techniques were devised to force the system to cross the barrier or to stay in the vicinity of a defined state by using biasing potentials [61]. We will detail the main ones used in this thesis in the next subsections.

1.3.5 Metadynamics

To overcome this limitation, several methods have been devised. One of them is metadynamics [2]. It was initially proposed by Alessandro Laio and Michele Parrinello and has been improved by several other additional tricks [5, 91] during the past 20 years.

To illustrate the working principle of this method, I like the example taken by Alessandro Laio and Francesco Gervasio in ref [3]. Classical unbiased MD can be seen as a walker falling into an empty pool during the night and exploring the bottom of it. The probability for the walker to go out of the pool is almost zero because the walls are too steep, so he will only explore the deepest point of the pool and its vicinity without exploring parts that are above. Now, if one gives the walker an infinite amount of sand, he will be able to deposit a small heap of sand at his current position. Then, little by little, the pool will start to fill in and at some point the walker will be able to escape the pool. This is the basic principle of metadynamics: a MD simulation is performed, and every t_d MD steps, a biasing hill is added to the potential energy surface of the form:

$$V_B(t, s) = \omega \sum_{k=1}^{n} \exp\left(-\frac{(s - s(kt_d))^2}{2\sigma_s^2}\right) \tag{1.54}$$

The free parameters of metadynamics are t_d, ω the height of the biasing potentials and σ_s the width of the hills. They are often chosen by looking at the variations of the CV in the basins.

If the metadynamics simulation is run for a long enough time, the biasing potential should compensate the form of the underlying free energy surface (FES) and hence, by keeping a history of the added Gaussians, at the end of the simulation one could compute the FES. It is also a great mean to obtain a first guess of the transition mechanism between the reactants and products. This is mostly how metadynamics is used in this thesis.

Nevertheless, we will present a small result on the Lennard-Jones dimer to conclude this example of the toy model. We performed a metadynamics simulation on the r_{dimer} coordinate, and collected the final free energy profile. Once the simulation is finished, the free energy profile is obtained by summing all the biases and taking the opposite. The results are presented in figure 1.4. At the end of the simulation, the system should evolve in a flat free energy landscape, which should result in transitions occurring instantaneously. This is very different from the start of the simulation, where no transition can happen. This can be seen on figure 1.4 (left) where at the start of the simulation the system cannot cross the barrier and stays in the equilibrium well, then by progressively adding the bias potential, the system can escape and recross the barrier several times to obtain a converged free energy profile figure 1.4 (right).

Another way of forcing the transition is by increasing the temperature, but this would also impact the mechanism of transition and also maybe destroy the useful information and the molecules. Because metadynamics is strongly dependent on the CV, the simulation might not evolve in the right direction. This is why, in this thesis, it was only used

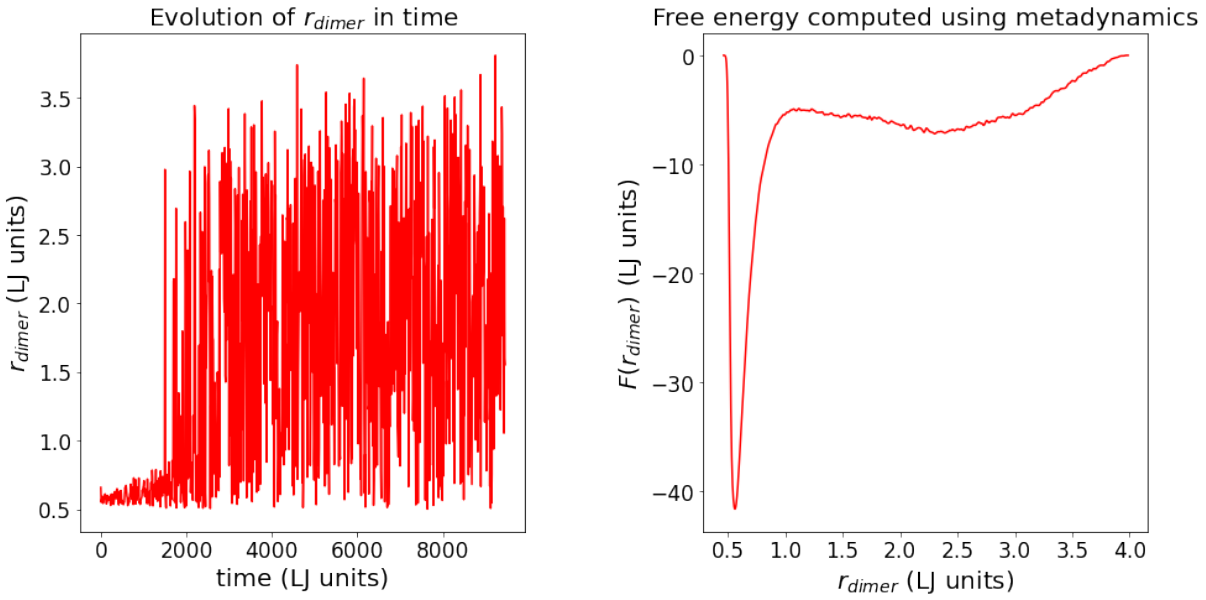

Figure 1.4: Metadynamics simulation of a Lennard-Jones dimer with interaction force of 40 LJ units, between the two particles of the dimer. Left, time behavior of the collective variable r_{dimer} during the simulation. Right: the resulting free energy along the r_{dimer} collective variable.

to obtain a first transition between the reactants and the products. To sample the free energy of the transition mechanism, we use Umbrella Sampling (US) [4]

1.3.6 Umbrella sampling

Once a first transition between the reactants and products is obtained and that we have first idea of the transition mechanism, we can extensively sample the CV-space to obtain the free energy surface using US.

When performing metadynamics simulations, the bias potential is progressively added during the simulation. Therefore, the potential energy surface on which the system evolves constantly changes. In umbrella sampling, the CV-space is cut into bins that we will call "windows". To sample different parts of the CV-space, quadratic potentials are introduced in each window and a single simulation is run. The resulting simulations are then unbiased using the so-called weighted histogram analysis method (WHAM) that we will present in the next paragraph.

The quadratic potential on CV s in window j centered on s_j is given by:

$$V_{bias,j}(s) = \frac{k}{2}(s - s_j)^2 \tag{1.55}$$

The key is to choose well the spring constant so that there is enough overlap between adjacent windows. The summary of the two enhanced sampling simulations presented in

this chapter is reported in figure 1.5. The figures are adapted from the thesis of Sara Laporte [58].

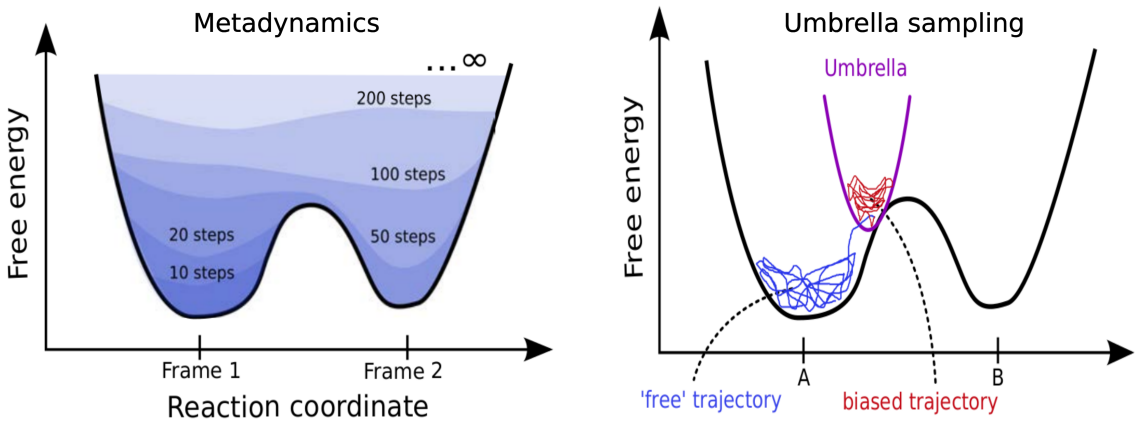

Figure 1.5: Summary of the two enhanced sampling methods presented in this chapter. Left: metadynamics, where a bias potential is progressively added in the potential energy surface making it a time dependent biasing method. Right: Umbrella sampling where several simulations are performed all along the chemical space under a quadratic bias potential. Figures are adapted from ref [58]

Unbiasing simulations with the weighted histogram analysis method

Once all the US simulations are performed, we get in each window the histograms of the collective variable under the biased potential, which is different for every window. Therefore, we need to put all the histograms on the same level, *i.e.*, we need to unbias the biased simulations. To do so, we use WHAM that is based on the WHAM equations that can be solved self-consistently just as the Hartree Fock equations or the Kohn-Sham equations. There exist many ways of deriving the WHAM equations [92, 93]: for example, one can derive these equations by minimizing the variance between all the windows of the same bin of the histogram [92]. But, since this thesis is oriented towards machine learning, we will adopt here a maximum-likelihood based approach that is akin to machine learning methods. We will sketch the basic principles of this proof based on likelihood and Bayesian approaches [94, 93, 95].

Let us suppose that N simulations are performed under the biasing potential given in equation 1.55. From each simulation j, we get a time series $(\mathbf{r_{i,j}})_{i\in[1,N_j]}$ of positions of the system that can then be projected onto the collective variable space: $(s_{i,j})_{i\in[1,N_j]}$, where N_j is the number of samples in simulation j. Each simulation can then be binned into M histograms: $n_{j,l}$ centered in s_l representing the number of counts in bin l. Finally, the normalizing condition is:

$$\sum_{l=1}^{M} n_{j,l} = N_j \tag{1.56}$$

From this data, we want to establish the true underlying probability p_l of having the CV in bin l. Given p_l, we can express in each biased simulation j, the probability of the

collective variable to be in bin l, $p_{j,l}$:

$$p_{j,l} = f_j c_{j,l} p_l \tag{1.57}$$

Where $c_{j,l}$ is given by the biasing potential at the center of the bin:

$$c_{j,l} = \exp\left(-\beta k (s_l - s_j)^2\right) \tag{1.58}$$

and f_j is a prefactor ensuring normalization for simulation j such that:

$$f_j = \frac{1}{\sum_{l=1}^{M} c_{j,l} p_l} \tag{1.59}$$

Given the probabilities $(p_{j,l})$ the likelihood of observing $(n_{j,l})$ in the respective bins is given by the multinomial law [95]:

$$L_j(n_{j,1}, \ldots, n_{j,M} | p_{j,1} \ldots p_{j,M}) = \frac{N_j!}{\prod_{l=1}^{M} (n_{j,l})!} \prod_{l=1}^{M} (p_{i,l})^{n_{i,l}} \tag{1.60}$$

And the total likelihood for the N US simulations to have collective variable in bin l is given by:

$$L((n_{j,l})_{j\in[1,N]} | p_1 \ldots p_M) = \prod_{j=1}^{N} L_j(n_{j,1}, \ldots, n_{j,M} | p_{j,1} \ldots p_{j,M}) \tag{1.61}$$

Now that we have the likelihood of observing our data given the probabilities p_l, we want to find the (p_l) that are most likely given the observations that we make, we will have a maximum likelihood approach in a Bayesian context. To do so, we use Bayes theorem that allows to reverse probabilities:

$$L(p_1 \ldots p_M | (n_{j,l})_{(j\in[1,N])}) = \frac{L((n_{j,l})_{j\in[1,N]} | p_1 \ldots p_M) P(p_1 \ldots p_l)}{P((n_{j,l})_{(j\in[1,N])})} \tag{1.62}$$

We will make the assumption that the prior distribution $P(p_1 \ldots p_l)$ is uniform, and therefore, when differentiating with respect to p_l this term will disappear, the term $P((n_{j,l})_{(j\in[1,N])})$ does not depend on p_l and therefore we can also discard it. Now, since differentiating a logarithm is easier than differentiating a product, we take the logarithm of equation 1.62 and combine it with equations 1.61, 1.60 and 1.57 to get the quantity to minimize:

$$B(p_1, \ldots, p_M) = \sum_{j=1}^{N} N_j \log f_j + \sum_{j=1}^{N} \sum_{l=1}^{M} n_{j,l} p_l + C \tag{1.63}$$

where C is a constant that contains all the terms that don't explicitly depend on p_l, we can differentiate this equation and get the final WHAM equation:

$$p_l = \frac{\sum_{j=1}^{N} n_{j,l}}{\sum_{j=1}^{N} N_j c_{j,l} f_j} \tag{1.64}$$

Together with equation 1.59, equation 1.64 forms the WHAM equations, they are coupled equations that need to be solved self-consistently as implemented in the Grossfield code [96] that we used in this thesis. One can use the Bayesian approach to assess the uncertainty of the WHAM procedure.

1.4 The sampling protocol: CV definition and applications to origins of life research

1.4.1 Introduction

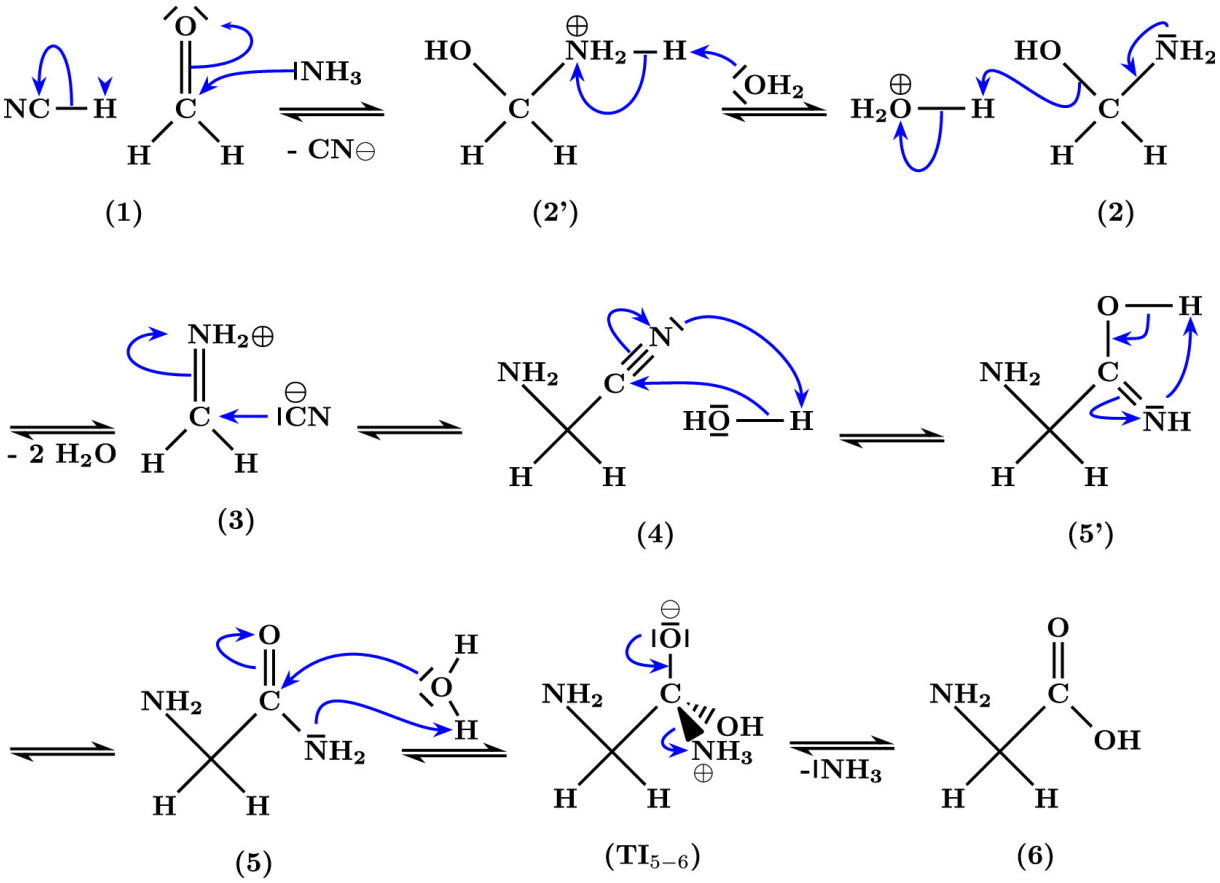

Figure 1.6: Strecker mechanism for the synthesis of glycine

Proteins are a key to all living organisms, they are constituted of a chain of amino-acids. There exists 22 different amino acids, the simplest of them is glycine. In prebiotic chemistry, the pathway often invoked for the synthesis of glycine is the Strecker [39] pathway. It is a mechanism from hydrogen cyanide, formaldehyde, and ammonia in water. Even though it is well known and cited, some steps of this pathway were not well-characterized, experimentally. This is why it was studied by a previous PhD student of the group: Theo Magrino [7]. Thanks to *ab initio* calculations, even short-lived intermediates can be identified, and the transition mechanisms can be observed. During his thesis, Theo Magrino set into place a whole protocol to sample chemical reactions in solution. In this section, we will present this protocol that was partly used in this thesis. The final mechanism is presented in figure 1.6.

1.4.2 The protocol

We will present the protocol devised in the team in order to have an agnostic exploration of the transition mechanism. With this protocol, only the reactants and the products need to be defined before performing the exploratory phase and then the sampling phase.

Path collective variables

First, as explained before, the exploration of a reaction mechanism and its sampling analysis relies on a CV. For the Lennard-Jones dimer, we used r_{dimer} as a CV because we knew we were studying a dissociation process. Nonetheless, here we are looking for a CV which does not rely on the prior knowledge of the transition mechanism and which can grasp the whole complexity of the transition mechanism.

We decided to use path collective variables (PCV) [5], they allow locating the desired configuration along a given chemical path given by the reference configurations $(X_\alpha)_{\alpha\in[1,N_{conf}]}$. This is done by computing the distance between the current configuration and the reference configurations in a metric space of dimension much lower than $3N$. They are formally defined in equation 1.65,

$$\begin{cases} s(t) = \left(\dfrac{\sum_{\alpha=1}^{N} \alpha \exp(-\lambda D[x(t), X_\alpha])}{\sum_{\alpha=1}^{N} \exp(-\lambda D[x(t), X_\alpha])} \right) \\ z(t) = \dfrac{-1}{\lambda} \log \sum_{\alpha=1}^{N} \exp(-\lambda D[x(t), X_\alpha]) \end{cases} . \tag{1.65}$$

The variable s measures the progress along the path given by the reference configurations, X_α while z characterizes the deviation at time t from these reference configurations. The key idea of these PCV is the first dimensionality reduction performed to go from the $3N$ dimensional space of atomic positions, to a lower dimensional space to compute the PCV.

We used a metric that allows to grasp the solvation effects of the reactive atoms [97], which is based on the computation of coordination numbers:

$$C_{i\sigma} = \sum_{j\in\sigma} \frac{1 - \left[\dfrac{r_{ij}}{r_0^{\alpha\sigma}}\right]^m}{1 - \left[\dfrac{r_{ij}}{r_0^{\alpha\sigma}}\right]^n} \tag{1.66}$$

This function computes how many atoms of atom type σ are within a radius $r_0^{\alpha\sigma}$ of atom i with atom type α. This function is computed for every reactive atom with respect to every atom type, resulting in a $N_{reac} \times N_{types}$ coordination table, where N_{reac} is the number of reactive atoms and N_{types} the number of different species in the system. If the parameters are well-chosen, this can account for covalent bonding, hydrogen bonding and long range interactions. The distance D is then computed using the following equation:

$$D(X_1, X_2) = \sum_i \sum_\sigma (C_{i\sigma}(X_1) - C_{i\sigma}(X_2))^2 \tag{1.67}$$

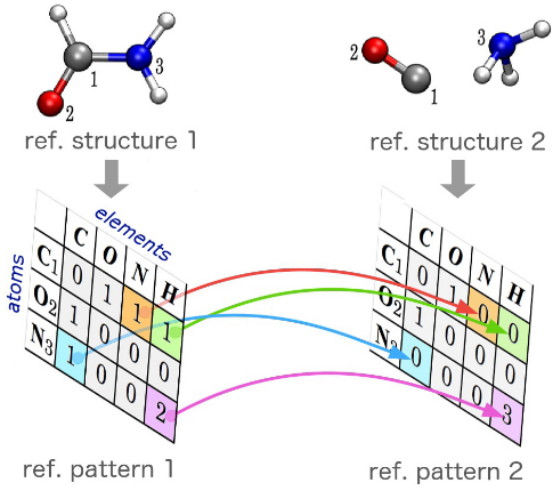

Figure 1.7: Representation of the PCV with the coordination table as a metric, adapted from [97]

The summary of the use of coordination table as a metric space is found on figure 1.7 and adapted from reference [97]. The key is now to find the given reference configurations.

Reference configurations

First, when we want to study a chemical transformation, the reactants and products are defined, simulation boxes are prepared in each state. Then, the mean of the coordination of the reactive molecules are computed and put as reference states.

Therefore, now PCV s and z can be built with these two reference configurations, they will be denoted s_2 and z_2. A metadynamics simulation is then run on (s_2, z_2) to obtain a first transition trajectory between the reactants and products. From this first trajectory, we want to characterize the TS and hence find the top of the barrier. To do so, we select configurations that seem to us close to the TS and launch a dozen of Hamiltonian trajectories from these configurations with initial velocities chosen from Boltzmann distribution. If between 30% and 70% of the trajectories reach the reactants basins and the other part the products basins, the configuration is categorized as a TS. This procedure is called "committor analysis" because the committor is a quantity defined as the probability of a configuration to fall in the products basin.

After performing committor analysis, the trajectories of this analysis are taken and 10 configurations are chosen among the frames of these trajectories to build a more accurate set of PCV. The reference configurations are chosen according to a Nudged elastic band Monte-Carlo procedure so that in the end, the chosen reference frames are equidistant in the latent space of the coordination table. The summary of the protocol to sample a reaction mechanism is given in figure 1.8.

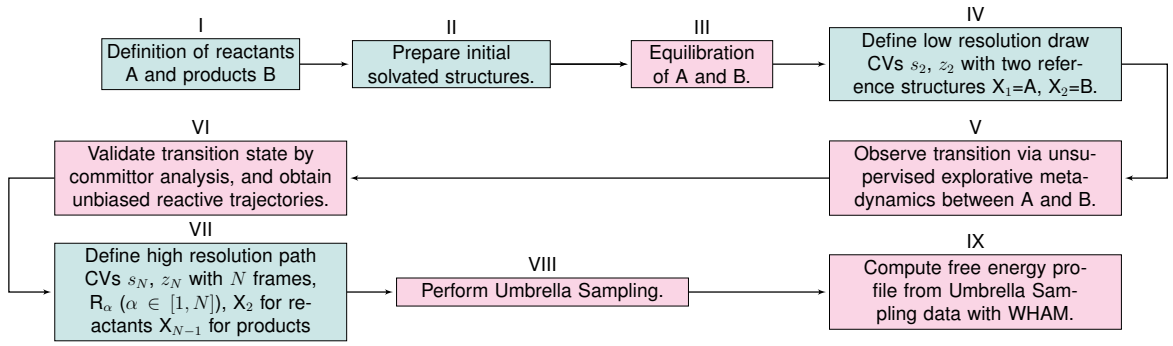

Figure 1.8: Schematic algorithm depicting the simulation protocol. Grey blocks (I, II, IV, VII) indicate pre- / post-processing steps where no simulations are needed. Red blocks (III, V, VI, VIII, IX) indicate agnostic explorative steps and expansive sampling steps performed using *ab initio* molecular dynamics. Adapted from [7]

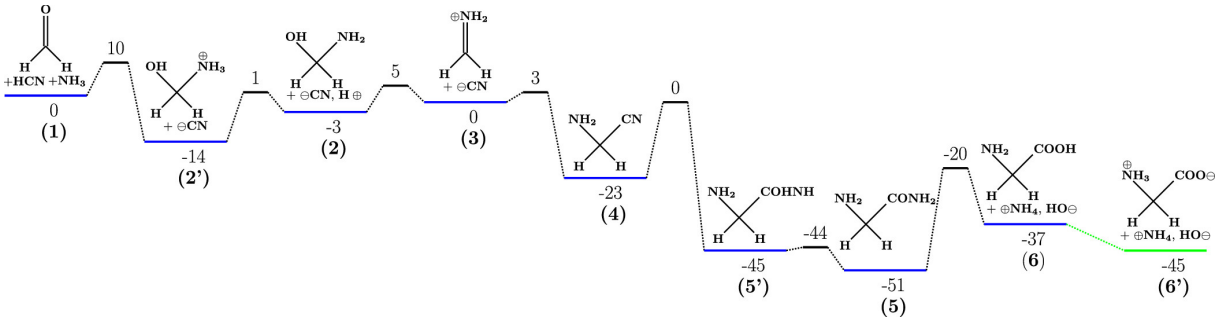

Figure 1.9: Free energy diagram of the whole Strecker synthesis of glycine

1.4.3 Results

The results of the simulations performed by Théo Magrino can be found in figure 1.9. From a box with water, hydrogen cyanide, ammonia and formaldehyde he characterized all the intermediates to go to glycine with their relative free energies using the protocol presented in the previous section. The fact that the amino-nitrile is very stable explains the observations of meteorites, but also the fact that this intermediate is observed, and no glycine is observed in the interstellar medium. This dataset generated by Theo Magrino was used in this thesis as a base dataset for machine learning purposes.

1.5 Partial conclusion

Throughout this chapter, we thoroughly explained all the methods used to study chemical reactions in solution using *ab initio* electronic structure methods to get the energies and forces. We then presented how we can simulate the dynamics of these reactions using molecular dynamics techniques and the energies and forces computed using electronic structure methods. One of the bottleneck however is that the covalent bonds are too strong to be broken during one MD simulation, therefore, one has to use enhanced sampling methods such as metadynamics or umbrella sampling to force the system to leave the equilibrium basins. We showed that these methods can be used to sample complex reaction

networks, such as the 7-step prebiotic Strecker-cyanohydrin synthesis of glycine. On the other hand, one bottleneck of such studies is the computational time. Indeed, the complexity of a DFT calculation grows with the cube of the number of valence electrons, thus, only small systems containing at most a thousand of atoms can be studied for at most 1ns. To overcome this problem, machine learning methods have been devised where the energies and the forces are learned so that the electronic structure calculations are replaced by a machine learning model. In the next chapter, we will present the basic principles of machine learning and how they can be used in atomistic simulations.

Chapter 2

Machine learning and its application to *ab initio* molecular dynamics

In this chapter, we will present the basic principles of machine learning. Then we will show how machine learning techniques can be applied to atomistic simulations.

2.1 Supervised machine learning

Machine learning (ML) and artificial intelligence are part of our everyday life. Even though, machine learning is included in artificial intelligence, we will here only talk about machine learning. There exists two types of machine learning:

- Supervised learning: we have a data set of features $X = (\mathbf{x_1}, \dots \mathbf{x_N})$, $\mathbf{x_i} \in A$ with their associated features $Y = (\mathbf{y_1}, \dots \mathbf{y_N})$, $\mathbf{y_i} \in B$ and we want to learn the relation between $\mathbf{x_i}$ and $\mathbf{y_i}$, *i.e.*, we want to find a function f such that $f(\mathbf{x_i}) \approx \mathbf{y_i}$. The difficulty of this task relies on finding the appropriate functional form for f in order for it to approximate the best the data.
- Unsupervised learning: we have a data set of features $X = (\mathbf{x_1}, \dots \mathbf{x_N})$, $\mathbf{x_i} \in B$ and we try to find underlying properties of the dataset, for example, can we find clusters in these data?

In this thesis we will only be interested in the first kind of approach, the supervised learning one. To summarize this approach I like the point of view of a member of the group, Arthur France-Lanord: in physics we have underlying models, for example Maxwell's equations in electromagnetism, we have an input that is the environment and the initial conditions, and we put them in Maxwell's equations to have the electromagnetic behavior of the system. When doing machine learning, we have the input (environment and initial conditions), the output (the electromagnetic field), and we try to find the underlying rules that link the output to the input.

2.1.1 Basic principles of machine learning

In this section, we will present the basic ideas behind supervised ML through a simple example of a noisy function that we want to fit. Let us assume that our features are

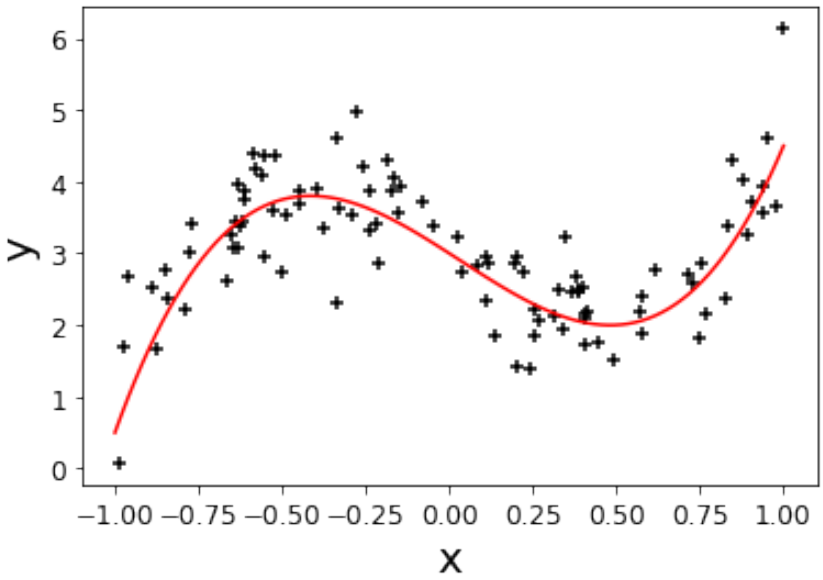

Figure 2.1: Example of a fitting situation where a function is sampled with a stochastic noise

deterministic and that only the labels are stochastic, *i.e.*, the labels can be separated as a sum of a deterministic function of x: g and a stochastic part ξ:

$$y_i = g(x_i) + \xi \tag{2.1}$$

with ξ some random variable on which we don't have to make any assumption for now. The typical situation is illustrated in figure 2.1. As explained before, our goal is to find the best function to approximate Y, the theoretical limit for this is naturally g, however, in practice, g is not known. We will use a maximum likelihood approach to know how we can infer these data. As done in the derivation of the WHAM equations, the aim is to derive a likelihood of the data given a model, and then, using Bayes theorem, to compute the probability that the model is likely given the data. The model we want to optimize is parametrized by a set of weights $\mathbf{w}$, the probability of having x_i given the model is:

$$p(\mathbf{x_i}, \mathbf{y_i}|\mathbf{w}) = \int p(\mathbf{x_i}, \mathbf{y_i}, \xi|\mathbf{w}) d\xi = \int p(\mathbf{x_i}, \mathbf{y_i}|\xi, \mathbf{w}) p(\xi) d\xi \tag{2.2}$$

Now, given ξ and $\mathbf{w}$, it follows:

$$p(\mathbf{x_i}, \mathbf{y_i}|\xi, \mathbf{w}) = \delta(\xi - (y_i - f_{\mathbf{w}}(x_i))) \tag{2.3}$$

Which yields to:

$$p(\mathbf{x_i}, \mathbf{y_i}|\mathbf{w}) = \mathbb{E}_\xi \left[\delta(\xi - (y_i - f_{\mathbf{w}}(x_i)))\right] \tag{2.4}$$

To process further we need to make assumptions on the distribution of ξ, as there are many y_i it seems natural to make the following Gaussian ansatz:

$$p(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\xi^2}{2\sigma^2}\right) \tag{2.5}$$

Hence:

$$p(\mathbf{x_i}, \mathbf{y_i}|\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f_{\mathbf{w}}(\mathbf{x_i}))^2}{2\sigma^2}\right) \tag{2.6}$$

Which allows us to write the likelihood of the dataset:

$$L(\mathbf{X}, \mathbf{Y}|\mathbf{w}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f_{\mathbf{w}}(\mathbf{x_i}))^2}{2\sigma^2}\right) \tag{2.7}$$

Using Bayes theorem, the posterior distribution of the parameters can be obtained:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) \propto L(\mathbf{X}, \mathbf{Y}|\mathbf{w}) p(\mathbf{w}) \tag{2.8}$$

with $p(\mathbf{w})$ the prior distribution of the parameters. Now we will start with a prior distribution that is uniform. We want to find $\hat{\mathbf{w}}$ such that the previous probability is minimized with respect to $\mathbf{w}$. As we did for the WHAM derivation, we take the logarithm to simplify the derivations. Therefore, the problem that needs to be solved is:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(y_i - f_{\mathbf{w}}(\mathbf{x_i})\right)^2\right] = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{w}|\mathbf{X}, \mathbf{Y}) \tag{2.9}$$

We find here a key result, that links the Bayesian analysis to a standard minimization problem. Indeed, we could have introduced *ad-hoc* the empirical risk that we wanted to minimize, because here it is the same as minimizing the square distance between the data and the model. The function that needs to be minimized, $\mathcal{L}(\mathbf{w}|\mathbf{X}, \mathbf{Y})$, is called the loss function. In the next section, we will see the prototypical examples of machine learning applied to linear models.

2.1.2 Properly training a model: example with polynomials

Up until now, we have been very general, our aim for this section is to show the basic principles of machine learning for a concrete case of regression on a polynomial function that has been sampled with some noise, see figure 2.1.

The situation is the following: we made a series of measurements for values of x between -1 and 1 $(x_1, \ldots, x_N)$ of a noisy signal $y(x)$ $(y_1, \ldots, y_N)$, and we want to find a law for the behavior of y as a function of x. The first thing we have to do is to specify the functional form of $f_{\mathbf{w}}$. For now, we will consider a polynomial ansatz:

$$f_{\mathbf{w}}(x_i) = \sum_{k=0}^{M} w_k x_i^k = \mathbf{w}\mathbf{x_i} \tag{2.10}$$

with:

$$\mathbf{x_i} = (1, x_i, x_i^2, x_i^3, \ldots, x_i^M) \tag{2.11}$$

The minimization problem of equation 2.9 can then be written:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left(\mathbf{Y} - \mathbf{X}\mathbf{w}\right)^2 \tag{2.12}$$

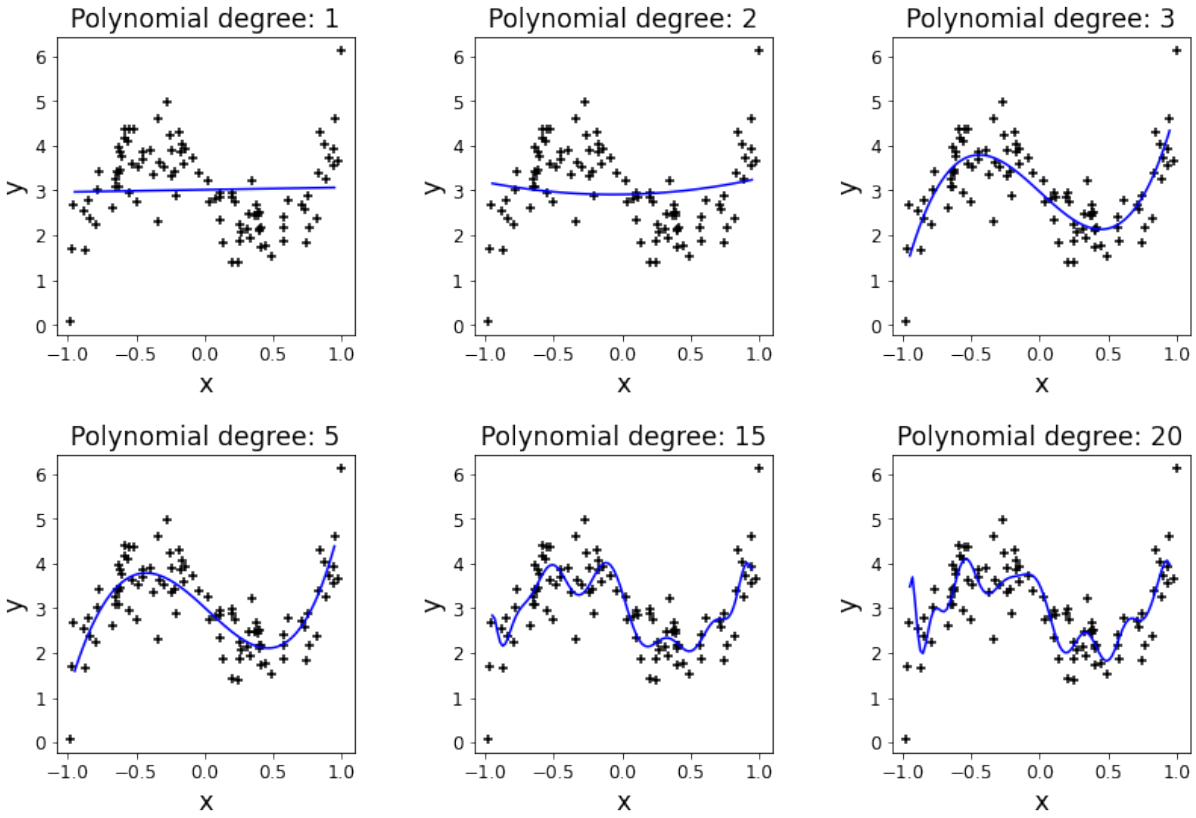

Figure 2.2: Examples of the fitted polynomial (blue lines) with respect to the measured signal (black dots)

with: $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ and $\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^M \\ 1 & x_2 & x_2^2 & \dots & x_2^M \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^M \end{pmatrix}$ $\mathbf{X}$ is called the design matrix.

Equation 2.12 can be solved very easily by differentiating it, the solution is:

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.13}$$

In ML, parameters that impact the efficiency of the model that are not parameters of the model itself such as the weights and that can be tuned are called "hyperparameters". They are nonetheless directly impacting the accuracy but also the complexity of the model. For example, in this very simple case, there is only one hyperparameter that is the degree of the polynomial function that we want to use, the higher the degree, the more complex our model will be. Let us see what happens when we increase it.

The results are shown in figure 2.2, we start by applying equation 2.13 with a polynomial of degree 1, still, the number of weights of the model is too low to grasp the complexity of the function that we want to fit. By increasing the degree of the polynomial, and thus the complexity of the model, we manage to reproduce the behavior of the measured data (degrees 3 and 5). On the other hand, if the degree of the polynomial is too high compared to the complexity of the measured data, the model is complex enough

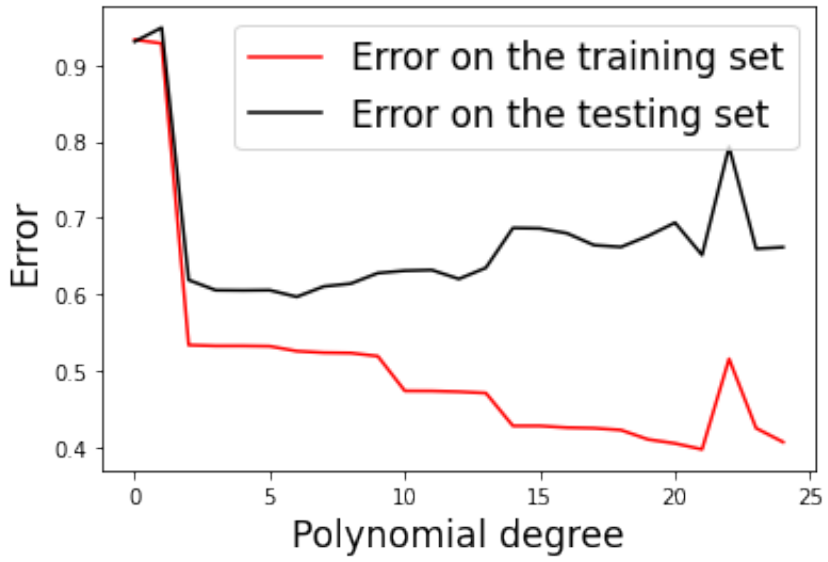

Figure 2.3: Testing and training set error as a function of the degree of the polynomial fitted on the training set

to adjust to every measured point and will therefore lead to what is called overfitting: the model is very good when predicting points that are included in the design matrix and that were used to perform the regression, but it will perform very poorly when predicting points outside the design matrix.

To have a more quantitative analysis of this problem in ML, the set of measured data can be split into two subsets: the training set which will be used to build the model, here in the case of the polynomial regression, only the points in the training set will be used to solve equation 2.12. The second set will be used as a test-set to assess the generalization error of the model: can the model we have trained be good at predicting data it has not seen.

This is illustrated in figure 2.3 where the training error only decreases when the complexity of the model increases, while the testing error increases when the model becomes too complex for the data. In a practical case, one has to find the point where the testing set error stops decreasing and starts increasing, this is the optimal value for the training of the model. This is done for example, by using cross validation.

By introducing a penalty term on the weights of the model in the loss function during the training, one can avoid overfitting. This is done in ridge regression, where the optimization problem is written:

$$\mathcal{L}(\mathbf{w}|\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{N} \left(y_i - f_{\mathbf{w}}(\mathbf{x_i})\right)^2 + \lambda \sum_{k=1}^{M} w_k^2 \tag{2.14}$$

This minimization process can actually be linked to the Bayesian approach we had previously. Indeed, following on equation 2.8, we chose a uniform prior distribution of

weights, but if instead, a Gaussian distribution is chosen, one gets the loss function of equation 2.14. Now, if we choose a Laplace distribution for the prior distribution of the weights ($p(\mathbf{w}) \propto e^{-|\mathbf{w}|}$), we get the following loss function:

$$\mathcal{L}(\mathbf{w}|\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{N} \left(y_i - f_{\mathbf{w}}(\mathbf{x_i})\right)^2 + \lambda \sum_{k=1}^{M} |w_k| \tag{2.15}$$

This last optimization problem is called the LASSO problem. When taking this problem from the prior probability point of view, the Laplace distribution is narrower than the Gaussian distribution. This is translated by the fact that he LASSO loss helps to identify useful degrees of freedom, for example by setting the useless weights to zero. All these processes are called regularization processes.

2.1.3 Gradient descent algorithms:

In the previous section, we had a closed form solution for our predictors. Moreover, the computational cost of inverting a matrix could be quite high if the dataset is high dimensional in features and in the number of data. To solve the optimization problem, gradient-based techniques have been put into place, where the solution is found by iteratively changing the weights towards the low gradient region. In this subsection, we will draw the basic principles of gradient descent methods.

We want to find a predictor $f_{\mathbf{w}}$ that minimizes a loss function given a dataset $(\mathbf{X},\mathbf{Y})$ $\mathcal{L}(\mathbf{w}|\mathbf{X},\mathbf{Y})$. To do so, we introduce the following sequence: starting with random initial weights $\mathbf{w^{(0)}}$, the following equation is iterated:

$$\mathbf{w^{(n+1)}} = \mathbf{w^{(n)}} - \eta_n \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w^{(n)}}|\mathbf{X},\mathbf{Y}) \tag{2.16}$$

Where η_n is called the learning rate. With this algorithm the weights are oriented towards the negative gradient and therefore should reach at some point a local minimum of the loss function. This nonetheless depends on the choice of the learning rate: if it is chosen too small, convergence will be slow. On the other hand, if it is too high, the algorithm will not be stable, and will only oscillate.

Often, the learning rate is varying during the gradient descent (hence the dependence in n), and decreases according to a power law or an exponential law. The reason for this is that often when performing optimization tasks, we start far from the global minimum of the function and therefore, we don't risk missing the minimum by going down very fast on the loss landscape. But as the algorithm progresses, we are getting closer to the minimum, this is why the learning rate decreases with n.

However, often in ML we are dealing with very complex landscapes, with many local minima, therefore, the gradient descent algorithm might lead to some local minimum and stay stuck in it. To escape a local minimum in physics, one often introduces stochastic thermal fluctuations. Moreover, the loss function is a sum over the whole dataset, which makes the evaluation of its gradient computationally expensive. To tackle these problems,

stochastic gradient descent algorithms have been put into place [98].

First, the dataset is randomly split into subsets that are called minibatches, the size of minibatches is called the batch size. At each iteration of the algorithm, the gradient will be computed only on one of the minibatches which is chosen randomly. This method adds stochasticity in the optimization process, which lowers the chances of getting stuck in a local minimum. It is also thought to bring natural regularization and hence prevents from overfitting [99, 100].

Nowadays, algorithms have been devised to keep track of the landscape on which they evolve by using momentum: if the landscape is flat and smooth, one will want to make big steps, while if it is steep and irregular, one will want to go slowly on the descent. For example, the ADAM algorithm [101], which is the one we will be using and is reported in algorithm 1. It uses first and second momenta to choose the orientation of the descent and to have an adaptive learning rate. In algorithm 1, β_i^n denotes β_i to the power n.

Algorithm 1 Adam optimizer for stochastic optimization of the loss function $\mathcal{L}$

Require: α: Step size
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $\mathcal{L}(\mathbf{w})$ stochastic loss function to optimize with respect to $\mathbf{w}$
Require: $\mathbf{w_0}$: Initial weight vector
  $\mathbf{m_0} \leftarrow 0$ Initialize first moment
  $\mathbf{v_0} \leftarrow 0$ Initialize second moment
  $n \leftarrow 0$
  while $\mathbf{w}^{(n)}$ not converged **do**
    $n \leftarrow n + 1$
    $\mathbf{g_n} \leftarrow \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w^{(n-1)}})$
    $\mathbf{m_n} \leftarrow \beta_1 \mathbf{m_{n-1}} + (1 - \beta_1)\mathbf{g_n}$
    $\mathbf{v_n} \leftarrow \beta_2 v_{n-1} + (1 - \beta_2)\mathbf{g_n^2}$
    $\hat{\mathbf{m}}_\mathbf{n} \leftarrow \mathbf{m_n}/(1 - \beta_1^n)$
    $\hat{\mathbf{v}}_\mathbf{n} \leftarrow \mathbf{v_n}/(1 - \beta_2^n)$
    $\mathbf{w^{(n)}} \leftarrow \mathbf{w^{(n-1)}} - \alpha \hat{\mathbf{m}}_\mathbf{n}/(\sqrt{\hat{\mathbf{v}}_\mathbf{n}} + \epsilon)$
  end while
  return $\mathbf{w^{(n)}}$

The first moment $\mathbf{m_n}$ and the second moment ($\mathbf{v_n}$) are computed in this algorithm. The second moment is used to set an effective step size in the stochastic gradient descent algorithm. We can see this in terms of variance, let us write ${\sigma_\mathbf{n}}^2 = \mathbf{v_n} - (\mathbf{m_n})^2$. The difference between the weight j at time $n-1$ and at time n, Δw_j can be written:

$$\Delta w_j = -\alpha \frac{m_n^j}{\sqrt{\sigma_n^j + m_n^j} + \epsilon} \tag{2.17}$$

Therefore, in case of small variance, $\Delta w_j \approx -\alpha$ while in the case of large variances : $\Delta w_j \approx -\alpha m_n/\sigma_n$. The learning rate is thus adaptive to the signal/noise ratio. This is what makes ADAM so powerful and so widely used.

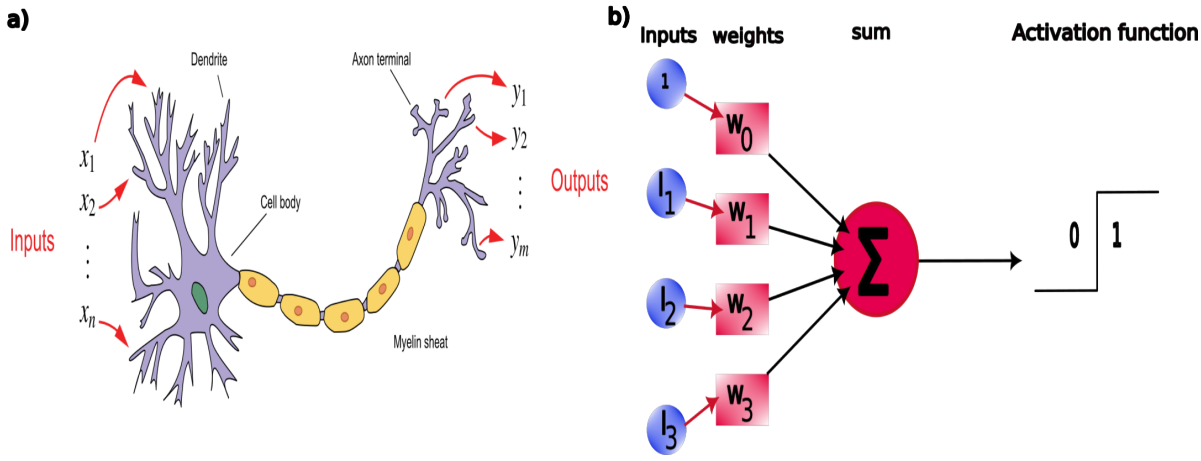

Figure 2.4: Panel **a)** the schematic representation of a neuron, and in panel **b)** its mathematical description by McCullogh and Pitts

2.1.4 Going non-linear with neural networks

Up until now, we presented the basic principles of ML using a very general formulation for the predictor ($f_{\mathbf{w}}$), or a very precise one in the case of the polynomial regression of the previous section. In this subsection, we will introduce a class of function that is often used in ML called neural networks.(NN) They allow having a very flexible predictor and are therefore suited for a wide range of problems.

Neural networks are the peak of the connectionism approach: from a set of inputs and outputs, we try to guess the underlying rules between them. With a neural network, this is done using a highly non-linear function. This gives a huge flexibility to the model. The first neural networks were devised by McCulloch and Pitts in 1943 [102] and were initially made to model a human neural network. The representation of a neuron and its mathematical modelling is shown in figure 2.4. The idea is that a biological neuron receives different inputs as an electrical signal, and depending on the values of the inputs gives 0 or 1 (signal or no signal) as an output. The model proposed is shown in figure 2.4, panel **b)**, inputs are now numbers, multiplied by weights; these products are summed and taken into what is called an activation function. Here the activation function is a step function, but in ML many functions can be used, such as the sigmoid, the hyperbolic tangent. These functions are a lot like the step function, but they have the advantage of being differentiable at every point. A single model neuron is called a perceptron and was first constructed by Rosenblatt in 1958 as a tool to identify images [103]. On his machine, inputs were the electric signals of photocells connected to potentiometers that act as the weights. The weights were tuned using electric motors.

The mathematical form of a perceptron is summarized by:

$$f_{\mathbf{w}}(\mathbf{x}) = \sigma\left(\sum_{i=1}^{M} w_i x_i + w_0\right) \tag{2.18}$$

Where σ is the activation function. Rosenblatt's perceptron made a lot of noise in the

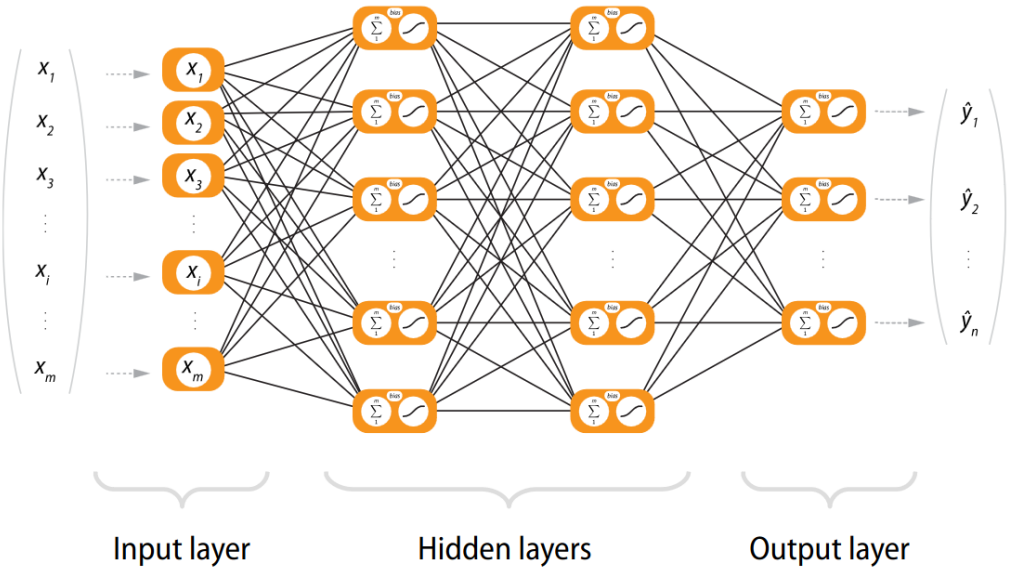

Figure 2.5: Schematic representation of the working principle of deep neural network, figure taken from a talk of Jean-Luc Parouty in [106]

community when it was created, because it was said to lead to "talking machines", but it was later shown that such a machine could only discriminate classes of pictures that were linearly separated [104]. This led to what is called the "first winter of artificial intelligence", but later on in 1986, a more complex form was proposed by a group of researchers in order to have more complexity in the model [105]. This is done by connecting perceptrons in layers, as presented in figure 2.5.

We can therefore have the following relation between layer l and layer $l-1$, if layer $l-1$ has n neurons and layer l has m neurons::

$$x_i^{(l)} = \sigma\left(\sum_{j=1}^{n} w_{ij}^{(l)} x_j^{(l-1)} + w_{i0}^{(l)}\right) = \sigma\left(z_i^{(l)}\right) \tag{2.19}$$

the weights w_{ij} can be put into the $m \times n$ weight matrix of layer l $\mathbf{W^{(l)}}$. The weights of this matrix need to be optimized. Such neural network is called a feed-forward neural network because the information is passed from a layer to another directly through neurons. We now want to optimize the matrix of weights according to a loss function $\mathcal{L}$. To do so, we want to use a stochastic gradient descent algorithm, as presented in the previous section. The gradient of the loss function is therefore needed, to compute it the backpropagation [105] method is used, since a neural network is just a composition of many functions. We want to evaluate :

$$\frac{\partial \mathcal{L}}{w_{ij}^{(l)}} \tag{2.20}$$

To do so, we will use the Leibniz chain rule of partial derivatives. After some algebra one

arrives to the following relation for the previous derivative:

$$\frac{\partial \mathcal{L}}{w_{ij}^{(l)}} = \frac{\partial \mathcal{L}}{\partial z_i^{(l)}} x_j^{(l-1)} = \Delta_i^l x_j^{(l-1)} \tag{2.21}$$

We now need to assess Δ_i^l, this once again can be done using the chain rule:

$$\Delta_i^l = \frac{\partial \mathcal{L}}{\partial z_i^{(l)}} = \sum_{k=1}^{m} \frac{\partial \mathcal{L}}{\partial z_i^{(l+1)}} \frac{dz_k^{(l+1)}}{dz_i^{(l)}} = \sum_{k=1}^{n} \Delta_k^{l+1} w_{ki} \sigma'(z_i^{(l)}) \tag{2.22}$$

We thus have equations for the gradient of each weight of layer l as a function of the gradient of layer l+1. These equations can thus be "backpropagated" to the first layer to get the gradients of the loss with respect to each weight. A stochastic gradient descent algorithm can then be used at the condition that we have an initial condition for the backpropagation. This is easily done by setting the last Δ_j^L as :

$$\Delta_j^L = \frac{\partial \mathcal{L}}{\partial z_i^{(L)}} \tag{2.23}$$

Which is easily computed given that the loss function is analytical.

2.1.5 The universal approximation theorem:

Neural networks are so widely used for a reason: they are thought to be universal approximators [107, 108]. Any "reasonable" function f can be approximated with an arbitrary accuracy by a neural network with some number of hidden layers, some number of neurons and some activation function σ. These theorems do not tell us the structure of the neural network as a function of the problem, nor how to optimize the weights. This theorem has a very simple proof. Let us assume the function f we want to approximate can be written as a primitive:

$$f(x) = \int_{-\infty}^{x} f'(y)dy \tag{2.24}$$

This can be written using the Heaviside function, and using approximate integration:

$$f(x) = \int_{\mathbb{R}} \theta(x-y) f'(y)dy \approx \sum_{i=-J}^{J} \theta(x - i\Delta x) f'(i\Delta x) \tag{2.25}$$

The Heaviside function can be approximated using a very steep sigmoid function σ:

$$f(x) \approx \sum_{i=-J}^{J} \sigma(\frac{x}{\epsilon} - \frac{i\Delta x}{\epsilon}) f'(i\Delta x) \tag{2.26}$$

Which can then be written:

$$f(x) \approx \sum_{i=-J}^{J} w_i \sigma(a_i x + b_i) \tag{2.27}$$

This is the structure of a single layer neural network. This means that any function can in theory be approximated to an arbitrary accuracy with a neural network. However, this theorem must be taken carefully, since it does not give any detail about the structure of the neural network, the weights, and the extrapolation accuracy.

2.2 Machine learning potentials

Now that we have seen the basic principles of machine learning and how to solve a machine learning problem, we will show how these methods can be applied to atomistic simulations and in particular AIMD simulations. Due to the fact that the Schrödinger equation has to be solved at each time step using DFT and that the force for each atom is needed, AIMD simulations are thus limited to the study of small systems (up to a thousand atoms and to the nanosecond).

Empirical force fields are a solution to this problem, but often they rely on a simple physically motivated functional form. This can lead to wrong results if the force-field is used for systems where the functional form is not appropriate.

This is why machine learning methods are brought in. The aim is to be able to predict the energies and the forces of a system, given the atomic positions, without solving the DFT equations. To do so, a sufficiently large dataset of configurations with their associated DFT energies and forces must be generated to train a machine learning model. In the next subsections, we will explain the architectures of such model and the physical motivations behind them.

2.2.1 Behler and Parrinello neural networks

The first neural networks potential for atomistic simulations were introduced in the 1990s with a very simple structure, yet they allowed to study realistic systems [109, 110, 111, 112]. But with the simplicity of the structure came also some drawbacks: first the structure of the neural network is fixed, the number of atoms cannot be changed. Moreover, the symmetries of the system were not embedded in the structure of the neural networks. Moreover, if two atoms of the same species are exchanged, the prediction of the NN will be different while it should stay invariant under permutations of atoms. To overcome these two problems, Behler and Parrinello in 2007 [9] came up with the simple idea to divide the total energy of the system as a sum of individual energies. Each individual energy being computed by a single neural network, *i.e.*, for a system with N atoms:

$$E_{tot} = \sum_{i=1}^{N} \epsilon_i \tag{2.28}$$

But, this still does not respect permutation invariance. To do so, the individual energies of the atoms of the same species are computed with neural networks having the same weights. This also has the effect of reducing the number of weights to optimize. The final

total energy of the system can thus be written:

$$E_{tot} = \sum_{i=1}^{N_{elem}} \sum_{j=1}^{N_i} \epsilon_i^j \tag{2.29}$$

Where N_{elem} is the number of different elements in the system and N_i is the number of atoms of element i. With this decomposition, a first training on a small system can be performed, and then more atoms can be added to perform MD simulations with the neural network potential (NNP). Nonetheless, if the input of the different NN are the Cartesian coordinates of the atoms the basic symmetries of the system are not respected: indeed, if a global translation is performed, all the coordinates change and the prediction of the NN too, but the system is the same. The same goes for global rotations of the system.

It is worth mentioning that although the total energy is divided as a sum of individual energies, each atomic energy taken separately does not have a physical meaning.

2.2.2 Respecting symmetries with atomic descriptors

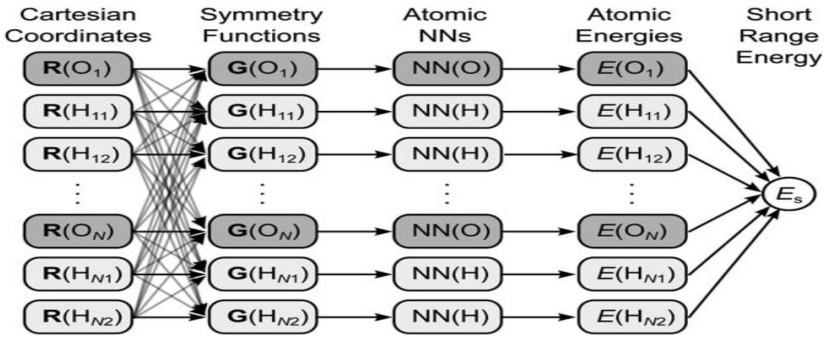

Figure 2.6: Schematic representation of a high dimensional neural network potential for a system of water molecules. The Cartesian coordinates of the system are transformed into sets of symmetry functions that are then given as input of atomic neural networks. The outputs of the atomic neural networks are then summed to give the final total energy.

To tackle this problem of symmetries, Behler and Parrinello came up with the general idea of descriptors. Instead of having the atomic coordinates as input of the NN, symmetry respecting functions of the system are computed for each atom and then given as input of the NN. They are therefore descriptors of the atomic environment of each atom. Thus, in many machine learning frameworks, the descriptors are computed within a cut-off sphere for every atom. In their seminal work, Behler and Parrinello introduced the "Atom centered symmetry functions" (ACSF) which are a set of functions based on radial and angular information to describe the environment of each atom. First, a cut-off function f_c is defined, it is a function of the interatomic distance which decreases smoothly from one to zero until the cutoff radius value. Then two sets of function are defined: the radial functions:

$$G_{i,\mu}^{rad} = \sum_{j=1}^{N_{atom} \in R_c} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \tag{2.30}$$

The summation is performed over all the neighboring atoms of the central atom. It is needed in order to keep the number of symmetry functions constant as an input of the NN. The η and R_s are a set of hyperparameters that needs to be chosen, there exists standard techniques to choose them as a function of the atomic species. This set of radial symmetry functions provide a good description of the radial information. The values of the sets of (η) and (R_s) can be automatically chosen according to automation procedures [113]. The second set of symmetry functions is used to encode the angular information between two atoms in the cutoff radius and the central atom in the following form:

$$G_{i,\mu}^{ang} = 2^{1-\zeta} \sum_{i,j,k} (1 + \lambda \cos\theta_{ijk})^{\zeta} e^{R_{ij}^2 + R_{ik}^2 + R_{jk}^2} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{2.31}$$

Where θ_{ijk} is the angle between the connections of the central atom i with neighbors j and k. As with the radial symmetry functions, the λ and ζ are sets of hyperparameters that can be chosen automatically. The final schematic representation of a neural network potential is shown in figure 2.6. Here we presented the historical approach to descriptors, but there exists a whole variety of descriptors that can be used in many situations [114, 115, 116].

From the definition of descriptors we gave, an analogy can be made with the concept of collective variables. Indeed, descriptors are just the projection of the environment of each atom onto a functional space. Moreover, descriptors can be used as a base of a data-driven CV built on a machine learning model [117].

2.2.3 The training process

As explained, the goal of a NNP is to be able to predict the energies of a system given the atomic configurations, but as shown in the previous chapter, to perform molecular dynamics simulations, one needs the forces, which are not directly given by the output of the NNP. But, since the neural network is a smooth differentiable function as well as the symmetry functions, the forces can be easily obtained with the chain rule. The force on the atomic coordinate α:

$$F_{\alpha,s} = -\frac{\partial E_{tot}}{\partial \alpha} = -\sum_{i=1}^{N_{atom}} \frac{\partial \epsilon_i}{\partial \alpha} = -\sum_{i=1}^{N_{atom}} \sum_{\mu=1}^{N_{sym,j}} \frac{\partial \epsilon_i}{\partial G_{i\mu}} \frac{\partial G_{i\mu}}{\partial \alpha} \tag{2.32}$$

As the end goal is to be able to perform molecular dynamics simulations, forces can be computed on the fly and taken as an objective in the loss function. The loss function to optimize is defined as follows:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{N_{struct}} \sum_{i=1}^{N_struct} \left[(E_{NN}^i - E_i)^2 + \frac{\beta}{3N_{atom}} \sum_{j=1}^{3N_{atom}} (F_{jNN}^i - F_j^i)^2 \right] \tag{2.33}$$

Where N_{struct} is the number of structures in the training set, E_{NN} the energy computed by the NN, E_i the reference energy of structure i, N_{atom} the number of atoms n in the system, F_{jNN} the j-th atomic coordinate of the forces computed by the NN, and F_j^i

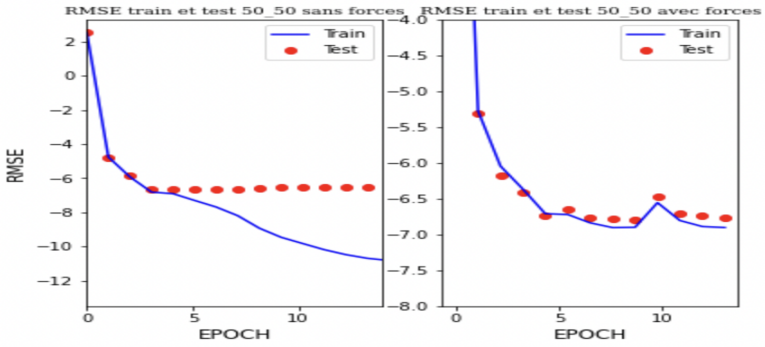

Figure 2.7: Example of training a neural network potential with only the energies in the loss function (left) and by including the forces in the loss function (right). Taken from Eliane Farhi's internship.

the reference one. β is a parameter to balance between the energy training and the forces training.

Using the forces not only allows adding them as an objective in the training, but also prevents from overfitting. This is shown in figure 2.7 that was obtained during the M1 internship of Elian Farhi that I supervised. The aim was to train a machine learning potential (MLP) for the study of CO2 under high pressures based on the data of M. Moog's thesis [118, 119]. It is clearly shown that in the case where the NNP is trained with only the energies, there is overfitting, while, when adding the forces in the training with the same dataset and the same structure, there is no trace of overfitting.

2.3 The example of the deepmd kit

The deepmd-kit [120] is a ready-to-use package for molecular dynamics simulations with machine learning potentials. The pre-processing of the data, generation of descriptors, training of the model is embedded in the kit. Then the trained model can be used to perform molecular dynamics simulations with popular codes such as LAMMPS, Gromacs, I-PI and so forth. The aim of the deepmd-kit, especially with the smooth edition, is to provide a tool close to *ab initio* methods where almost only the atomic coordinates and the atomic species need to be entered. In this section, we will explain how it works.

2.3.1 Descriptor computation:

First, the environment of each atom i is embedded in a matrix written $\mathcal{R}^i$ containing the vectors $\mathbf{r_j} - \mathbf{r_i}$ with $\mathbf{r_j}$ all the position vectors of the atoms of index j within a cutoff sphere of radius r_c centered on r_i. The components of $\mathcal{R}^i$ are then transformed into reduced coordinates $\tilde{\mathcal{R}}^i$, with the j-th row being: $(s(r_{ji}), \hat{x_{ji}}, \hat{y_{ji}}, \hat{e_{ji}})$, where:

$$\hat{x_{ji}} = \frac{s(r_{ji})x_{ji}}{r_{ji}} \tag{2.34}$$

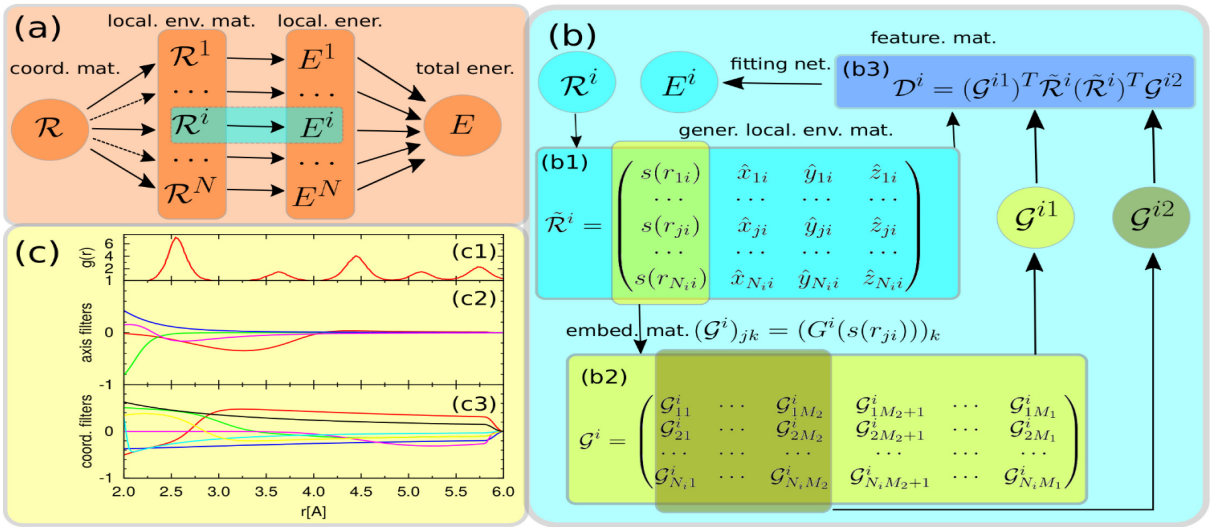

Figure 2.8: Summary of the working principle of the deepmd package, smooth edition. Adapted from [121]

And the same is done for $\hat{y_{ji}}$ and $\hat{z_{ji}}$. $s(r_{ji})$ is a smoothing function defined as follows:

$$s(r_{ji}) = \begin{cases} \dfrac{1}{r_{ji}} & r_{ji} < r_{cs} \\ \dfrac{1}{r_{ji}}\left[\dfrac{1}{2}\cos\left(\pi\dfrac{r_{ji}-r_{cs}}{r_c-r_{cs}}\right)+\dfrac{1}{2}\right] & r_{cs} < r_{ji} < r_c \\ 0 & r_{ji} > r_c \end{cases} \tag{2.35}$$

It allows having components to go smoothly to zero in the same fashion as the cutoff function of the symmetry functions seen in the previous section. This decrease is controlled by the "smooth-cutoff" parameter r_{cs}. It can be shown that the product:

$$\Omega^i = \tilde{\mathcal{R}}^i(\tilde{\mathcal{R}}^i)^T \tag{2.36}$$

is invariant under rotations and translations, it is however not invariant under permutations. To make it so, a new matrix is introduced:

$$\mathcal{D}^i = (\mathcal{G}^{i_1})^T\tilde{\mathcal{R}}^i(\tilde{\mathcal{R}}^i)^T\mathcal{G}^{i_2} \tag{2.37}$$

The matrices $\mathcal{G}$ are called embedding matrices and are defined as follows:

$$(\mathcal{G}^i)_{jk} = (G(s(r_{ji}))_k \tag{2.38}$$

Where G is the local embedding network mapping a single input to M_1 outputs (resp M_2). $\mathcal{D}^i$ is thus a $M_1 \times M_2$ matrix that is flattened into a vector to be given as input of the atomic energy network.

2.3.2 Training a model with deepmd

The energy and the forces are computed using a neural network with a Behler-Parrinello structure. The loss function commonly used is the following:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{|\mathcal{B}} \sum_{l \in \mathcal{B}} p_\epsilon |E_l - E_l^{\mathbf{w}}|^2 + p_f |F_l - F_l^{\mathbf{w}}|^2 \tag{2.39}$$

It is optimized using the ADAM optimizer presented in the previous section. $\mathcal{B}$ is the batch, E_l and F_l are the reference energies and forces, while $E_l^{\mathbf{w}}$ and $F_l^{\mathbf{w}}$ are the ones computed by the neural network. p_f and p_ϵ are prefactors chosen arbitrarily.

Deepmd has shown great results on public dataset and on extending the size of the studied system to millions of atoms [122]. But new models built on graph neural networks are starting to outperform it [123, 124]. Deepmd-kit is more and more used to study realistic systems. We used it throughout this thesis on different typical reactions in solution. In the next chapters, we will show how we used deepmd to create a suitable training set to perform umbrella sampling simulations.

Part II

Application: Machine learning potentials for prebiotic chemistry

Chapter 3

Development of machine leaning potentials for chemical reactions in solution

3.1 Introduction

In this chapter, we present the results published during my PhD on a methodology to build a training set [16]. In the past few years, a number of machine learning methods and frameworks have been developed to tackle the problem of obtaining *ab initio*-level potential energy surface (PES) at a reasonable computational cost [8, 125, 9, 126, 120, 10, 127]. Although these methods have significant differences, the basic principle is common: using AIMD trajectories, a model is trained and then used to infer *ab initio* quality PES, and thus perform *ab initio* quality molecular dynamics simulations for a much lesser computational cost.

This approach can be used to access larger system sizes [122, 13], and/or to perform simulations longer than the time reachable using traditional AIMD simulations, in order to observe interesting physical transformations and/or improve statistical sampling [128].

However, two important questions still need to be fully addressed in ML-based molecular dynamics: how to build optimal training sets, and how to critically assess the quality of machine-learning potentials in chemical reactions, a difficult setting where the system is led to explore very energetic configurations far from the geometries of the metastable minima.

An interesting tool in this respect is the "neural-network committee" method[129]. It consists in training several NNPs on the same training set but with different random seeds. In this way, for the same configuration, the NNPs will give different results. The standard deviation of the predictions of the different members of the committee on some observable is used to assess the reliability of the average prediction. Indeed a good agreement between the different NNP means the configuration is close-enough to the training set for the NNP to be accurate, while a higher value means that the prediction cannot be trusted.

This technique can be used to build a training set using an iterative training [113, 121] procedure. First, a small set of NNPs is trained with existing AIMD data; then, a MD simulation is run using one of the NNP of this set. After this step, configurations that display a standard deviation over the set of NNP on the prediction of some observable above a certain threshold, are recomputed using single point *ab initio* calculations and added to the training set. This is repeated until no more configurations are evaluated as mis-predicted.

The iterative learning framework has led to thorough studies of systems with *ab initio* quality at reduced computational cost[130, 131]. More recently, the committee method has been also used to quantify the error on an observable computed using NNP-based molecular dynamics simulations [132], as well as a way to iteratively select configurations from an AIMD trajectory to build optimal training sets [133]. A recent in-depth study of committee methods[134] has shown that in iterative schemes randomly selecting additional configurations to be evaluated at DFT level to improve the training set (random sampling) is equivalent to a selection based on committee disagreement beyond a threshold, but that the latter has to be carefully calibrated.

In this thesis we adopt NNP-driven enhanced-sampling molecular dynamics simulations to study chemical reactions in solution. This is a challenging goal as it requires the NNP to explore high-energy configurations far from equilibrium, with highly distorted chemical bond geometries. In particular, for a given $A \rightarrow B$ reaction, we aim at extensively sampling along the RC connecting the two basins, with the aim of reconstructing the accurate free energy landscape through US simulations. To this end, it is crucial to train NNPs capable to yield locally-accurate and well-behaved PES throughout the relevant reaction space.

Although enhanced sampling has been combined with machine-learning potentials in a few recent studies [135, 14, 132], including a combination of US with NNP[136], the critical assessment and systematic use of NNP for chemical reactions in solution are still lacking.

In this chapter we present benchmarks and construction principles for training sets. We carefully assess the total computational cost of the training and data production trajectories, with the goal of limiting the total amount of *ab initio* calculations without losing accuracy. We also introduce a simple approach to ensure long stable trajectories with high NNP-committee agreement: at variance with a previous method where the error is evaluated from deviations between DFT and NNP predictions,[15] our simple scheme avoids the burden of additional *ab initio* calculations.

We apply the new scheme to two reaction steps of the Strecker-cyanohydrin synthesis of glycine in water, previously studied at the DFT level[7]. Our results indicate that a surprisingly reduced amount of suitably-chosen *ab initio* samples is sufficient to train a NNP potential able to sample the full reaction coordinate space between minima, leading

to accurate free-energy profiles and barriers at a significantly reduced cost compared to purely DFT simulations.

3.2 The simulation setup:

3.2.1 Benchmark reaction

We first consider the initial step of the Strecker-cyanohydrin synthesis of glycine, particularly emblematic in prebiotic chemistry studies [1, 54]:

$$HCN + COH_2 + NH_3 \rightarrow CH_2OH(NH_3)^+ + CN^-$$

as illustrated in Fig. 3.1. The CVs regarding this reaction will be noted with a subscript $(1) \rightarrow (2')$[7]. The aim is to assess the behavior of a NNP-system, trained on a minimal amount of AIMD US trajectories, carefully selected along an optimized reaction pathway.

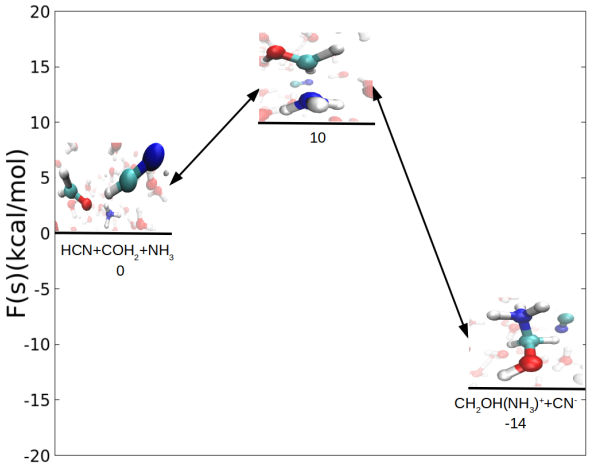

Figure 3.1: Free energy diagram of the first step of the Strecker-cyanohydrin synthesis of glycine obtained in our previous work [7]. The error is of the order of the kcal/mol and was estimated using block averages. The reactants, products, and transition state configurations are also reported on the graph.

3.2.2 Neural networks details

Since our target system is quite heterogeneous, including 251 atoms in a 13.4 Å-side cubic box under periodic boundary conditions, with the following composition: 2 C, 2 N, 6 H and 1 O atom in the reactive molecules, solvated with 81 explicit water molecules, whose H and O atoms are indistinct from the ones in the reactive part of the system, and treated exactly on the same footing by the NN. As a consequence, the water molecules would comparatively acquire an overwhelming weight, with respect to the important reactive subsystem, during training. To overcome this potential difficulty, we defined

a loss function (equation 3.1) capable to take into account this heterogeneity:

$$L(\mathbf{w}) = \frac{1}{|B|}\sum_{l\in B}\left[p_E|E_l - E_l^{\mathbf{w}}|^2 + p_f\frac{1}{N_{elem}}\sum_{i=1}^{N_{elem}}\frac{N_{atoms}}{n_i}|\mathbf{F}_l - \mathbf{F}_l^{\mathbf{w}}|^2\right] \tag{3.1}$$

,

where n_i is the number of atoms of type i in the system, N_{elem} is the number of different elements in the system, N_{atoms} is the total number of atoms in the system, E_i and F_i denote the DFT energies and forces of the training set, while $E_l^{\mathbf{w}}$ and $F_l^{\mathbf{w}}$ are the forces computed by the NNP, and B is the batch size.

This choice of the loss function allows weighting equally each element type via the error on forces. The weight p_f progressively increase from 1 to 10 while the opposite happens for p_e during the training, according to the protocol implemented in the deepmd-kit package [120].

The neural network used to compute the descriptors was made with 3 layers of 25, 50 and 100 nodes while the energies networks were made with three layers of 240 nodes each. All the AIMD configurations come from a previous study of the group[7]. We used LAMMPS [137] for all the NNP-based molecular dynamics simulations, employing hydrogen atoms instead of deuterium in the original study to have a quicker, more realistic dynamics. All the enhanced sampling simulations performed with neural networks were carried out using the open-source, community-developed PLUMED library [138], version 2.5.0 [139] .

3.2.3 Umbrella sampling set-up

We adopt a set of one-dimensional, quadratic umbrella sampling potentials applied on the s variable and centred at positions s_j expressed as:

$$V_{bias,j}(s) = \frac{k}{2}(s - s_j)^2 \tag{3.2}$$

Windows are equally spaced by $s_{j+1} - s_j = \Delta s$ based on $k = k_BT/(\Delta s/2.5)^2$ in order to have sufficient overlap between two windows (see Ref. [7]). The s path-CV measures the progress along a given reaction pathway, the sampling is therefore performed on that coordinate. The z path-CV measures the deviation from the pathway and helps to detect possible anomalies in the sampling. Once a specific reaction pathway is determined and targeted, US is performed on it, in order to determine the corresponding free-energy profile. Due to the intrinsically high-energy, unstable character of configurations explored close to the barrier top, in US simulations of chemical reactions it is occasionally observed that the system can deviate from the targeted pathway to explore a different one.

To focus the sampling on the reaction mechanism under study, a restraining potential is sometimes applied along z, as in the case of the step (1) $\rightarrow$ (2′) of reference [7]. The free energy profile is computed using the WHAM method presented previously [140] implemented in Grossfield's code [96]. We use a convergence criterion of 10^{-7} kcal/mol

and 150 bins in s space. We estimate the statistical uncertainty as the deviation between the free energy profiles computed using the third and fourth quarters of each trajectory. The obtained free energy profile is then used to compute the activation barrier by taking the free energy difference between the highest point in the profile and the reactants free energy; the free energy difference between reactants and products can also be computed. The value of the activation barriers can then be compared to experimental results.

A correction term can also be added to this free energy difference as it was done in reference [141]. However, since the scope of the present work is to compare NNP results to the AIMD ones of the original study [7], this additional term was not in this work.

From a ML point of view, the splitting of the RC space into independent US windows simplifies the selection of different data sets to form the training set, and the evaluation of the performance of the NNP.

3.3 Building the training set

3.3.1 Detecting the frontiers of accurate predictions in a neural network potential

The performance of a NNP is customarily measured via the error with respect to DFT on a training set and a testing set. However, this is not sufficient with respect to our goal, *i.e.*, accurately computing the full free energy landscape. In this respect, it is important to assess the capacity to generate long, stable and accurate MD trajectories. We base our assessment on the committee approach [129], including a time-dependent metric similarly to a recent study [15].

As shown in figure 3.2, when, during a simulation, the NNP-generated trajectory exits the "safe" region of configuration space, where forces are accurately predicted, the value of the z pathCV significantly increases, indicating a large deviation from the reference transition pathway and suggesting the likelihood of unphysical configurations. This is confirmed by the analysis of the corresponding structural properties of the system. An inspection of the C-O pair correlation functions $g(r)$ (figure 3.2 c)), before and after this jump, reveals that an unphysical short-distance peak has appeared, not present in the *ab initio* data. This confirms that the NNP sampled an unphysical region and is trapped in it, as the predicted energy decreases (figure 3.2 b), a spurious stable configuration. On the other hand, this transition corresponds to a sudden jump in the standard deviation in energy predictions among committee members as well as in the maximum standard deviation on the predicted atomic forces, defined in the following way:

$$\sigma_{max} = \max_{j \in [1, N_{atoms}]} \sqrt{\left[\sum_{i=1}^{4} \| \mathbf{F}_i^j - \mathbf{F}_{avg}^j \|^2 \right]} \tag{3.3}$$

where $\mathbf{F}_i^j$ is the force on atom j computed by NNP i, and $\mathbf{F}_{avg}^j$ is the average over all the neural networks. [14, 130, 131]. We remark that σ_{max} is a more general and accurate

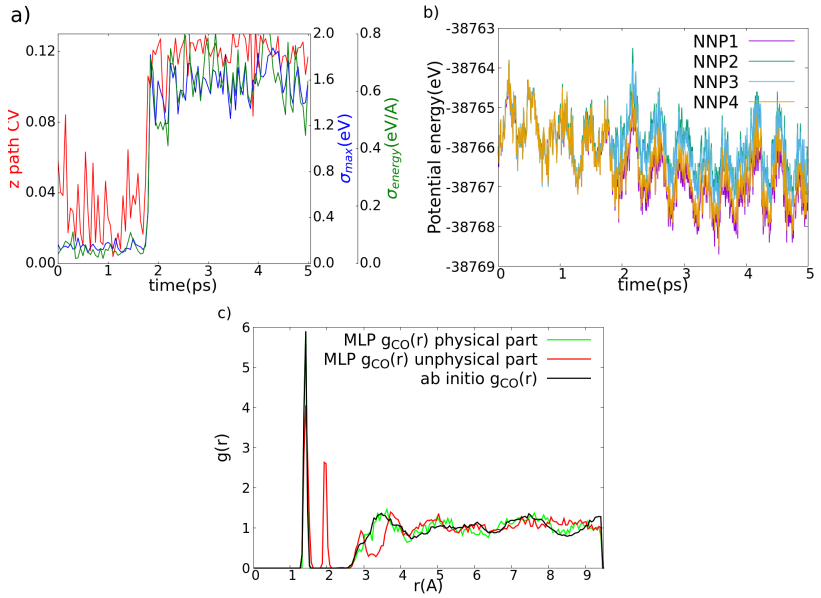

Figure 3.2: Assessing the quality of predictions of neural network potentials using a committee method. Panel **a)**: left (red line): time evolution of z path CV; right: time evolution of the maximum standard deviation on the prediction of forces(σ_{max}, blue line) and standard deviation on the prediction of energies(σ_{energy}, green line). Panel **b)**: time evolution of the potential energy predicted by the four neural networks. Panel **c)**: Carbon-Oxygen radial distribution function for the AIMD simulation (black line) and for the NNP simulation, before (green line) and after (red line) the unphysical jump in the predictions of energy and forces.

indicator of the loss of predictive power than the $g(r)$ or the pathCV z.

This approach allows detecting the frontiers of the configuration space region where the NNP gives accurate prediction. We can thus define a simulation lifetime τ as the time at which σ_{max} surpasses a reliability threshold, simply corresponding to the full simulation time if no pathological behaviors occur. All the trajectory before τ is physically sound, and can be employed to collect statistics. The next logical step is to devise a procedure to maximise τ all along the RC space as a function of the training set composition and extension, in order to carry out reliable US simulations.

3.3.2 Generating training sets suited for free energy calculations

In this section, we critically assess the effect of the composition and size of the training set on the error of the free energy profile recontructed with the NNP potential. The aim is to retain *ab initio* accuracy while minimizing the amount of DFT calculations necessary to train the potential.

Following the approach employed in our previous full-*ab initio* work[7] on the reaction considered here, we identify the following algorithm:

- Perform a preliminary DFT-based metadynamics simulation [2] employing pathCVs

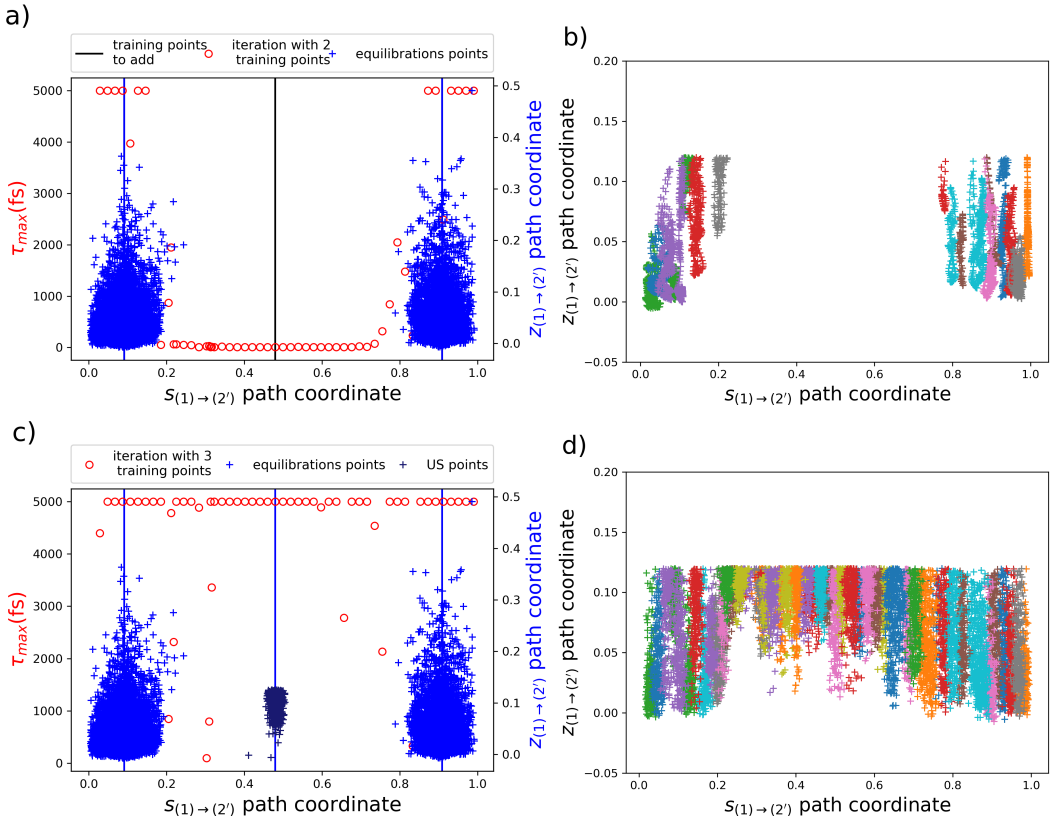

Figure 3.3: Building the training set of a neural network potential for a chemical reaction in solution. Panel **a)**: maximum lifetime τ_{max} (red dots) in each US window throughout the path CV for a NNP trained only on the AIMD equilibrations of the reactants and the products; the AIMD training points are represented in blue in the (s,z) plane. Panel **b)**: corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane; panel **c)** maximum lifetime τ_{max} (red dots) in each US window throughout the path CV for a NNP trained only on the AIMD equilibrations of the reactants, the products, and a transition state US window; the AIMD training points are represented in blue in the (s,z) plane. Panel **d)**: corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane

built upon the reactants and products as the only references[97]: this allows a prejudice-free exploration of a reactive pathway (quick, without the need to converge a free-energy estimate), to be refined through committor analysis and leading to the definition of improved pathCVs based on multiple reference structures (see Ref. [7] for details).

- Generate one AIMD US trajectory (of about 15 ps) in the transition state window along the optimized pathCV (defined at the previous step), and include it in the training set along with AIMD equilibration trajectories (of about 15 ps) of the reactants and the products. These three trajectories represent *ab initio* "milestones" of the RC to train and test the NNPs.

- Train four models on the same training set, with different random seeds. Define

a range of US windows densely spanning the full RC range, and for each window perform 50 short (about 5 ps) NNP-based US simulations, with the same starting point but different random initial velocities taken from the Boltzmann distribution.

- Plot the maximum lifetime τ_{max} over each US window (see figure 3.3): if in some RC region τ_{max} is smaller than the autocorrelation time of the pathCV, *i.e.*, if it is impossible to generate uncorrelated samples, it is necessary to generate an additional AIMD US trajectory in that region (since the NNP is locally unreliable) to be added to the training set.

- Repeat the two last steps until a satisfactory NNP is obtained, *i.e.* with an acceptable lifetime τ across the full RC space, hence capable of producing a dense sampling and a converged US free energy landscape.

The procedure is illustrated in figure 3.3, showing the training configurations for each iteration (about 600 structures saved every 20 fs) and the resulting NNP samples in the (s, z) plane and lifetime for every US window.

Clearly, adding training points according to our scheme allows to progressively increase the simulation lifetime and the overlap between NNP US simulations. In the next two sections, we discuss the error of the NNP on a testing set, and the accuracy of the NNP free energy landscape with respect to the *ab initio* one.

3.3.3 Error estimate on the benchmark reaction

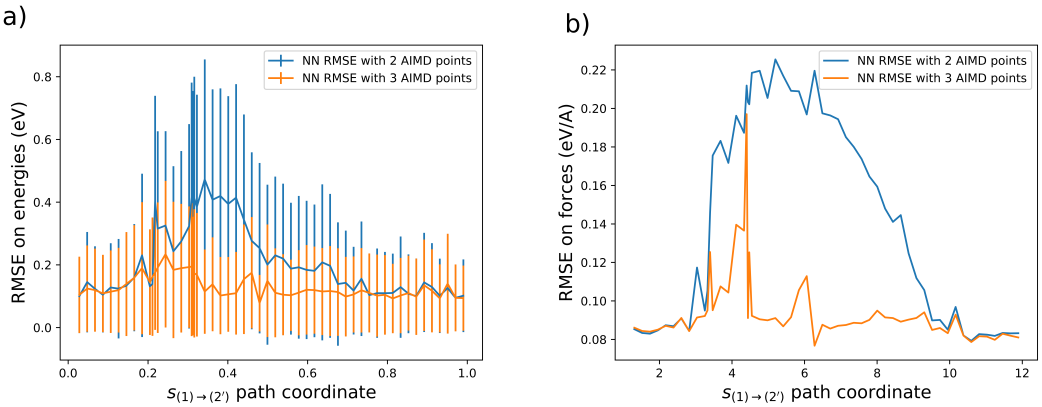

Figure 3.4: Root mean squared error on the energies and on the forces of a neural network potential along the chemical space. Panel **a)**: Root mean squared error (RMSE) of the NNP-estimated energy along the chemical path for a NNP trained only on the AIMD equilibrations of the reactants and the products (blue line) and for a NNP trained on those configurations and on the US transition state window (orange line). The test set is built only with configurations that are not present in any training set. Panel **b)**: same as panel **a)** but the RMSE was computed on forces instead of energies

Before proceeding on the actual calculation of the free energies, it is important to verify that the NNP displays a rather uniform error over the s-coordinate. To this aim,

we compute the root mean square error (RMSE) of the NNP-predicted energies and forces on a test set built with configurations from *ab initio* simulations. Those configurations were never included in any of the training sets. The results for the case of a training set including only reactants and products (2-AIMD points) are compared with the case including also one transition-state-like US window in the trainng set (3-AIMD points) in figure 3.4. Clearly, adding the third trajectory to the training set improves the error in the central region of RC space and leads to a uniform RMSE.

Although this error-evaluation step gives us an idea on the quality of the NNP produced using our new procedure, it cannot be applied on a system that has not been extensively studied with AIMD.

3.3.4 Calculation of the NNP free energy surface for the benchmark reaction

We now assess the accuracy of the free energy landscape reconstructed from the NNP. Figure 3.5 shows that it is possible to obtain a first-principle quality free energy surface using, in the training set, only one US AIMD simulations and the two end-point AIMD equilibration simulations of the reactants and the products, and to recover the full FES mostly via NNP-US windows. This is a major gain of computational time, which could allow to study, at the same level of *ab initio* accuracy, larger and more complex systems.

The full original *ab initio* FES in Ref. [7] was obtained using 55 US AIMD windows; our procedure significantly reduces, by more than an order of magnitude, the number of *ab initio* MD simulations to be carried out. From the computational point of view, the training of 1 neural network takes approximately 24 GPUh; while one 5-ps simulation takes about 0.05 GPUh. Hence, for the full study of this test reaction, we needed a total of 161 GPUh, to which one needs to add the 40k CPUh used to build the *ab initio* training set. The fact that all the calculations were run on GPU makes non-trivial the comparison with respect to the CPU simulations. However, the whole *ab initio* study of this reaction in Ref. [7] took around 700k CPUh.

In figure 5, the free energy was computed using short simulations of 5 ps per window. In the next section, we propose a method to generate longer, but still stable, NNP trajectories.

3.3.5 Generating stable NNP trajectories

After properly training our NNP, our aim is to be able to generate long and stable MD simulations. However, it is known that, at some point, the NNP-system will eventually explore untrained regions of the configuration space, where energies and forces will necessarily be extrapolated, likely failing to describe the correct system anymore. To overcome this common drawback, one possibility to restrain the sampling to low σ_{max} regions is to add a quadratic potential on σ_{max}, as it was proposed in reference [129]. In this work, instead, we find it more effective to completely avoid high-σ_{max} values, which would correspond to an infinite spring constant with respect to the above-mentioned reference.

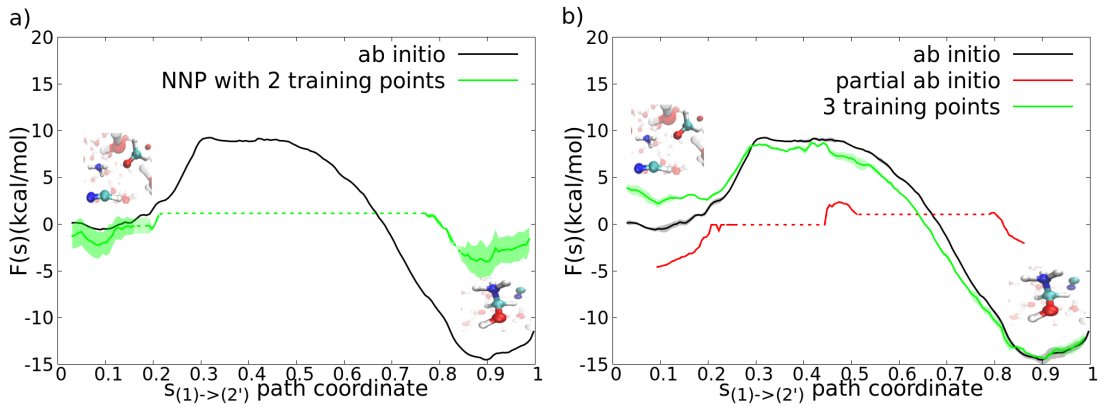

Figure 3.5: Free energies obtained using a neural network potential with an increasing training set size compared with the *ab initio* reference. Panel **a)**: US-obtained free energy profiles using full AIMD (black line), and using a NNP trained on AIMD equilibration of reactants and products (green line). The shaded zones correspond to the estimated statistical errors. Panel **b)**: US-obtained free energy profiles using full AIMD (black line), and using a NNP trained on AIMD equilibration of reactants, products, and the transition state US windows (green line).The red line represents the free energy which would be obtained using the three AIMD US windows (reactants, products, transition state). The reactants and products endpoint-configurations are also reported on the graphs.

Therefore, we can define such a "mirror reflection operation" as follows:

We identify two regions in the configuration space sampled by each US window: the region where the NNP behaves properly, *i.e.,* where the maximum standard deviation on the prediction of forces σ_{max} is low, and the region where significant deviations, hence

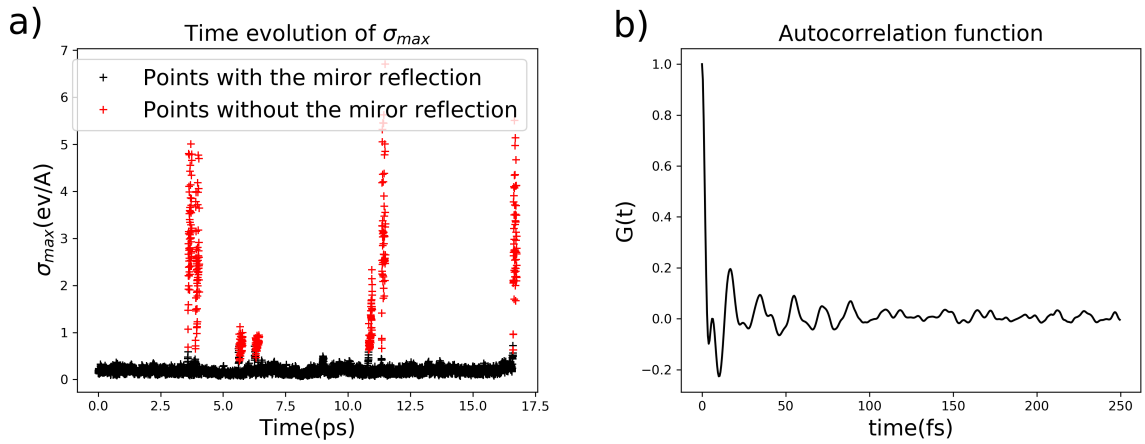

Figure 3.6: Panel **a)**: Time evolution of the maximum standard deviation on the prediction of forces (σ_{max}) during a NNP umbrella sampling simulation. We report with black crosses its instantaneous values before the non-physical behavior sets in, and with red crosses its behavior after that. Panel **b)**: NNP-autocorrelation function of the s reaction coordinate computed in the same NNP-umbrella sampling window.

nonphysical behavior, are observed.

Whenever σ_{max} exceeds a given threshold, we define for each atom α the following vector $\mathbf{G}_\alpha$, the value of this vector in the k direction is:

$$G_{\alpha k} = \frac{1}{n_{comm}} \sum_{i=1}^{n_{comm}} (F_{\alpha k}^{i} - F_{\alpha k}^{avg})^2 \tag{3.4}$$

With n_{com} the number of neural networks in the committee ($n_{comm} = 4$ in this study). This quantity provides the normal vector of a frontier between the reliable, "interpolation region" of the NNP, and a NNP-unknown "extrapolation region". We can therefore mirror-reflect the trajectory of the atoms on this hypersurface when σ_{max} displays a pathological jump to high values.

In order to do so, and force the NNP-system to remain in the interpolation region, the simulation is stopped a t_{step} number of time steps before the instability, and is restarted after imposing the following transformation on the velocities:

$$\mathbf{v}_\alpha^{new} = \mathbf{v}_\alpha^{old} - \frac{2\mathbf{v}_\alpha^{old}.\mathbf{G}_\alpha}{\|\mathbf{G}_\alpha\|^2}\mathbf{G}_\alpha \tag{3.5}$$

where $\mathbf{v}_\alpha^{old}$ is the velocity of atom α at the moment in which we mirror-reflect the trajectory, while $\mathbf{G}$ is evaluated at the moment in which where σ_{max} passes the instability threshold.

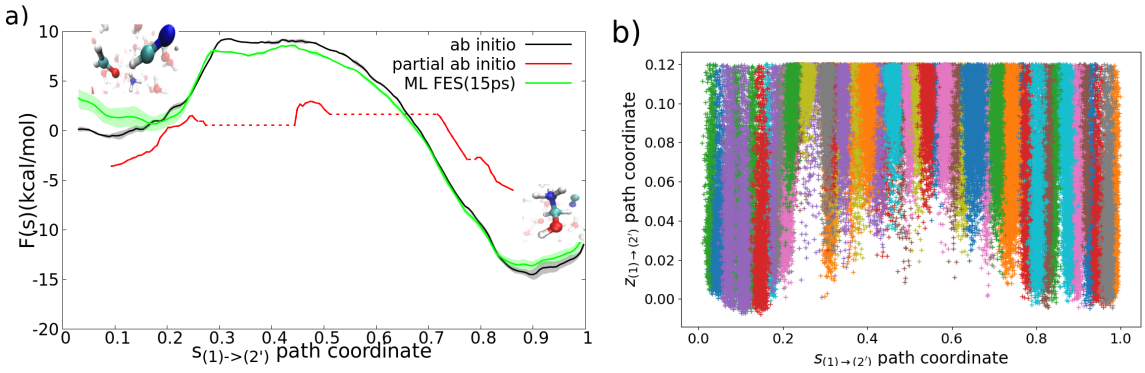

Figure 3.7: Panel **a)**: US-obtained free energy profiles using full AIMD (black line), and using NNP trained on AIMD equilibration of reactants, products and transition state US windows when the velocity transformation is applied (green line). The red line represents the free energy which would be obtained using only the three AIMD US windows (reactants, products, transition state). The reactants and products endpoint-configurations are also reported on the graph. Panel **b)**: corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane. As in the original study[7] z path CV was confined to within 0.12 via a semiparabolic wall.

This operation preserves the kinetic energy of the system, without pathological effects on the US procedure. We consider that it is a useful and legitimate complement to our NNP-based free-energy calculation protocol whenever the typical lifetime τ is significantly longer than the auto-correlation time of the CV, so that uncorrelated samples can be

	Ab initio study	machine learning study
$\Delta F^{\dagger}_{(1)\to(2')}$ (kcal/mol)	10	8 ± 0.7
$\Delta F_{(1)\to(2')}$ (kcal/mol)	-14	-14.5 ± 1

Table 3.1: Activation barrier ($\Delta F^{\dagger}$) and free energy difference between reactants and products($\Delta F_{(1)\to(2')}$) obtained in the *ab initio* study along with the ones obtained in this work

collected between successive reflections. In the opposite case, it is advisable to improve the NNP via a larger training set before employing it to perform statistical sampling.

To illustrate the choice of the parameters of the protocol, the time evolution of σ_{max} during a US simulation is reported in figure 3.6 along with the NNP-autocorrelation function of the CV.

The threshold of σ_{max} is chosen as 0.6eV/A, while choosing the number of time steps at which the reflection is performed before the instability between 200-500 gives similar results. Clearly, the typical time-lapse between two reflections (ranging here from 0.3 to 5 ps) is much larger than the autocorrelation time ($\sim$ 0.1 ps). Reflections therefore appear not harmful from a statistical viewpoint, and allow carrying out long, stable US simulations localized in the region of configuration space where the NNP is reliable, leading to free-energy estimations retaining *ab initio* accuracy (Figure 3.7 and table 3.3).

The profile obtained in figure 3.7 lies within the typical 2 kcal/mol uncertainty of the PBE functional for this kind of studies, as well as the predicted free energies, shown in table 3.3.

3.4 Application of the protocol to a more complex reaction

The aim of this part of our work is to challenge the method previously devised to reproduce a complex FES. To this end, we chose another intermediate step of the Strecker-cyanohydrin synthesis, presenting a more complex free energy landscape, consisting of a two-step process, as shown in figure 3.8: a hydrogen bond breaking between the imine nitrogen, a hydrogen atom and the cyanide, followed by the addition of the cyanide to the imine.

This process displays a small barrier of 5 kcal/mol, a small drop of 3 kcal/mol towards a metastable state, another small barrier of 3 kcal/mol with respect to the latter step, followed by a large 23 kcal/mol drop to the products, as obtained from our original AIMD FES in Ref. [7] by using 44 US windows. Such small free-energy barriers are likely more difficult to be quantitatively and qualitatively described by a NNP, than in the previous reaction. We underline that, although the stoichiometry of the chemical species is the same as in the previous case, the reactants and the products are different, and thus the NNPs are generated from scratch, independently from the previous case.

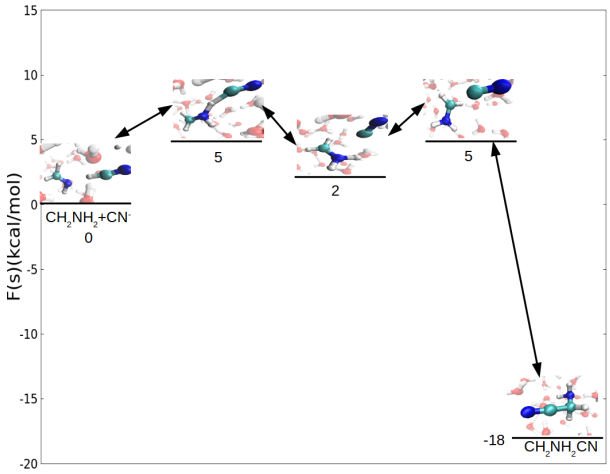

Figure 3.8: Free energy diagram of the reaction (3) → (4) of the Strecker-cyanohydrin mechanism[7]. The error is of the order of the kcal/mol and was estimated using block average [7]. The relevant configurations are also reported on the graph.

Results of US simulations employing a series of three NNPs with different training sets are presented in figure 3.9. We progressively enlarged the training set as previously described, by using at first 1, then 3 and finally 6 AIMD US windows, besides the reactants and products ones (respectively 3, 5 and 8 training points).

The NNP FES obtained starting from the different training sets are shown in figure 3.10. Judging from the accuracy of the FES, NNP training is optimal when 6 *AIMD* US windows are used, which is reasonable considering the fact that this reaction consists of two successive steps. Instead, unsurprisingly, the FES obtained with only 3 training points compares very poorly with the *ab initio* one.

Adding AIMD US windows to the training set leads to a progressive increase of the accuracy of the NNP FES, until a satisfactory agreement with the benchmark AIMD one is achieved. As in the case of the previous reaction, the cumulative duration of the AIMD trajectories necessary for training the NNP is one order of magnitude shorter than the total duration required for computing an accurate AIMD FES, despite the fact that the present reaction is clearly more challenging to be reproduced in its fine details. The free energy profile obtained using the velocity transformation technique is shown in figure 3.11. We report in Table 3.2 the CPU time comparison between a full ab initio FES calculation and a NNP-based one. As we have no guarantee that our NNPs are transferable to another reaction, we chose to add the CPU time needed to perform the training simulations in the comparison with the *ab initio* study.

3.5 Conclusions

In this work, we perform a critical assessment of different procedures to build a NNP training set starting from AIMD US trajectories of chemical reactions in solution, exploiting

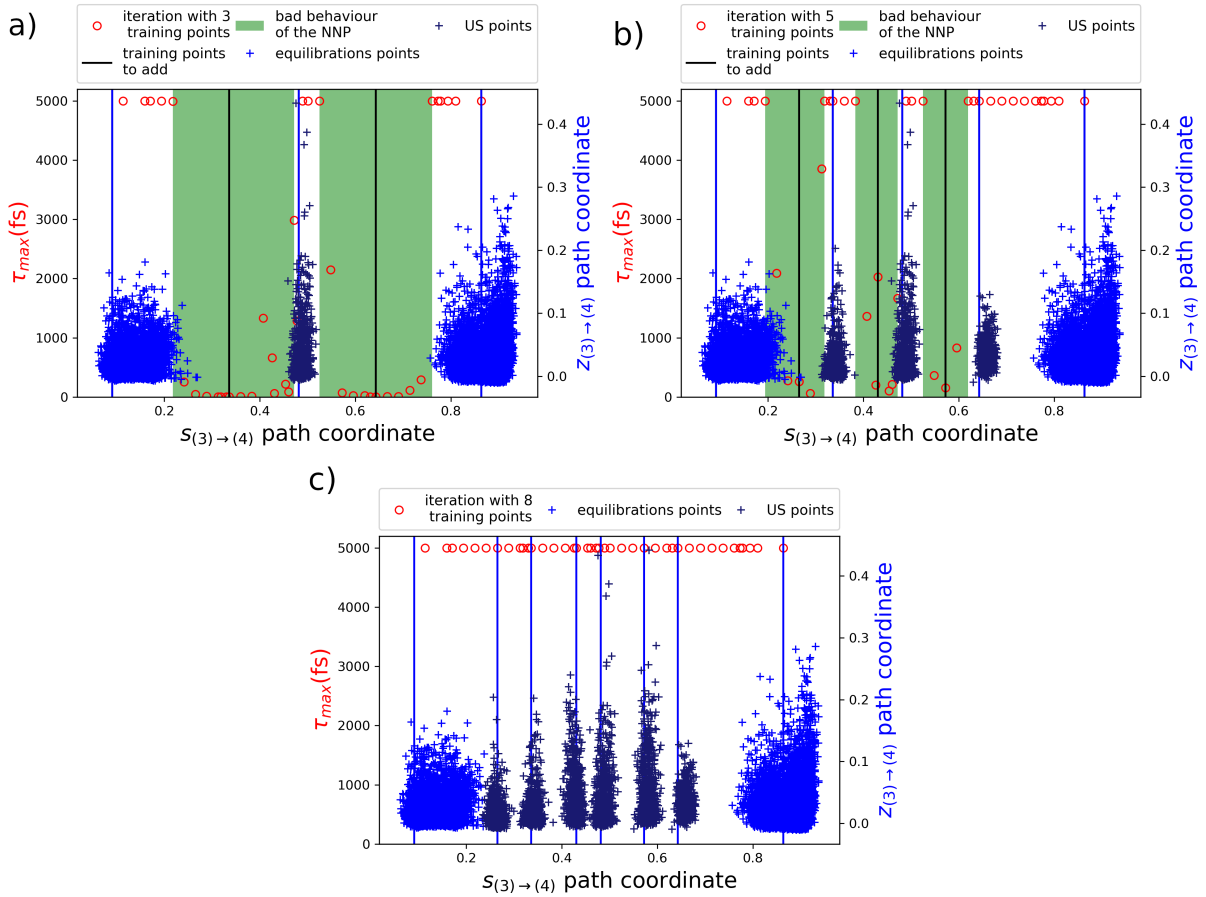

Figure 3.9: Illustration of the iterative procedure followed to build a converged NNP for reaction (3) → (4). Panel **a)**: maximum lifetime τ_{max} (red dots) in each US window throughout the path CV for a NNP trained only on the AIMD equilibrations of the reactants, the products, and the transition state; the AIMD training points are represented with blue crosses in the (s,z) plane. Panel **b)**: same as **a)** but adding two additional AIMD US windows in the NNP training set. Panel **c)**: same as **b)** but adding three additional AIMD US windows in the NNP training set.

two reactions along the Strecker pathway [7] as test cases. The systematic comparison of AIMD training sets increasingly more dense along the reaction coordinate clearly indicates a threshold for achieving NNP free energy profiles of ab initio accuracy.

We also provide an approach to exploit the disagreement between predictions of equally-trained NNPs not only as diagnostics but also to render robust and stable the long-time dynamics, hence to achieve satisfactory statistical sampling, a crucial advantage for free-energy reconstruction.

Our investigation suggests a new protocol for the accurate, ab initio quality, NNP calculation of FES of chemical reactions in solution: the computational load of pure and costly AIMD simulations would be limited to preliminary metadynamics exploration with mechanism-agnostic general-purpose CVs [97], followed by committor analysis to define mechanism-specific CVs [7], and finishing with a limited number of AIMD US trajectories along the reaction path, used for training a NNP.

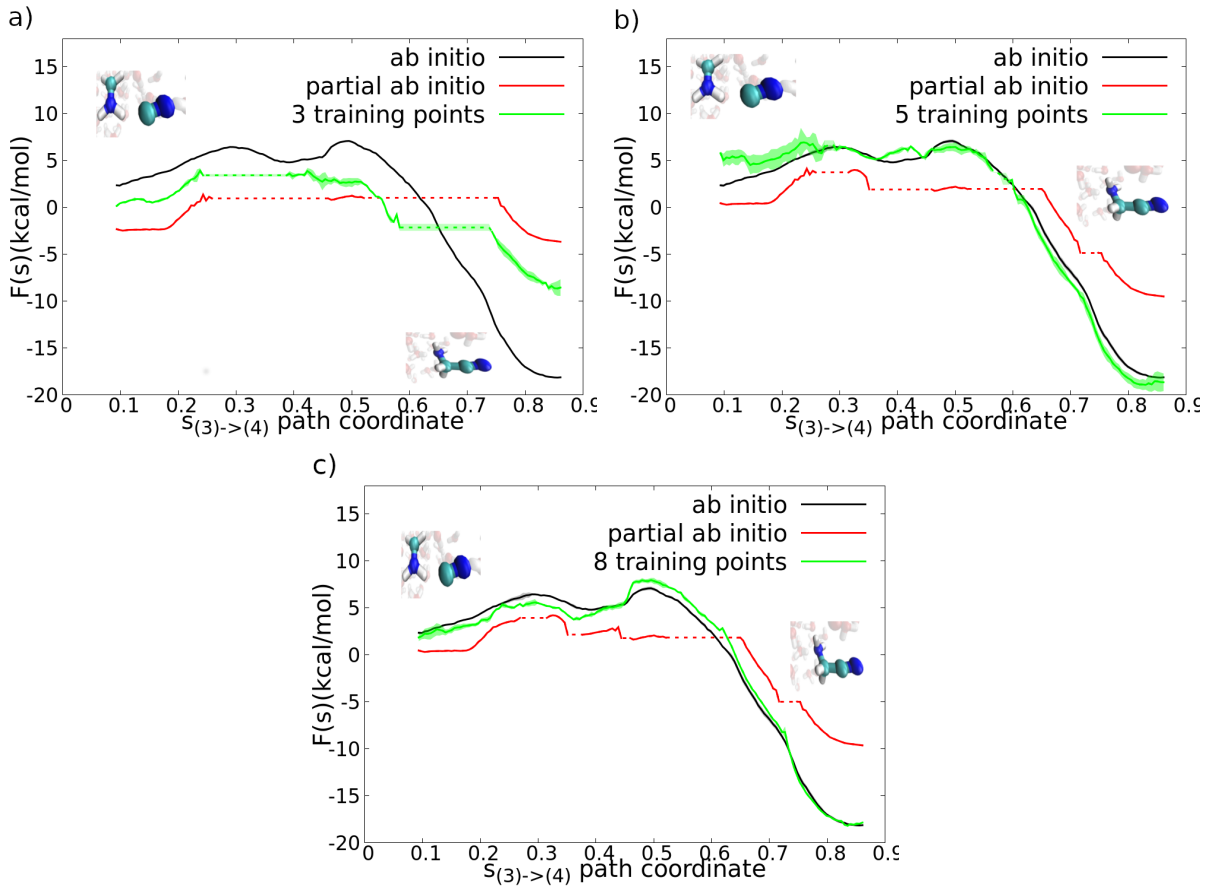

Figure 3.10: Free energies obtained for the reaction (3)→(4) of the Strecker synthesis with an increasing size of the training set. Panel **a)**: US-obtained free energy profiles using full AIMD (black line), and NNP trained on AIMD equilibration of reactants, products and transition state windows (green line). The red line represents the free energy which would be obtained using only the three AIMD US windows (reactants, products, transition state). The shaded zones correspond to the estimated statistical errors. Panel **b)**: same as **a)** but using the five training points in panel **b)** of figure 3.9. Panel **c)**: same as **b)** but using the eight training points in panel **c)** of figure 3.9. The reactants and products endpoint-configurations are also reported on the graphs.

	Ab initio study	machine learning study
training time	0	$24h(GPU) \times 4 \times 3$ iterations
US ab initio time	$15 \times 44 = 660ps$	$15 \times 6 = 90ps$
CPU/GPU simulation time	540k CPU.h	100k CPU.h + 200 GPU.h
Total CPU/GPU time	540kCPU.h	488 GPU.h + 100k CPU.h

Table 3.2: Summary and comparison of the computational times for the (3) → (4) reaction between the pure AIMD protocol and the combined AIMD-ML one.

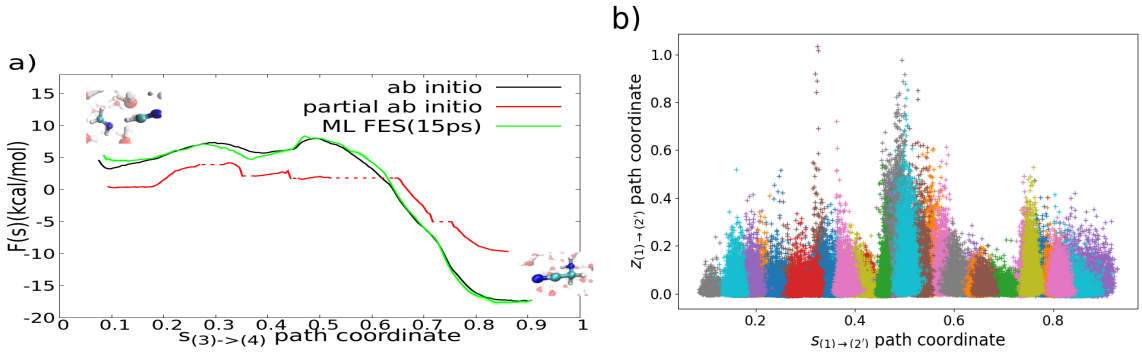

Figure 3.11: Free energy obtained with a neural network potential for the step (3) → (4) of the Strecker synthesis of glycine. Panel **a)** free energy profile obtained as in panel **c)** of figure 10 but applying our velocity transformation (equation (3.5)). Panel **b)** corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane.

	Ab initio study	machine learning study
$\Delta F^{\dagger}_{(3)\to(4)}$ (kcal/mol)	3	3.5
$\Delta F_{(3)\to(4)}$ (kcal/mol)	−20	−20

Table 3.3: Activation barrier ($\Delta F^{\dagger}_{(3)\to(4)}$) and free energy difference between reactants and products($\Delta F_{(3)\to(4)}$) obtained in the *ab initio* study along with the ones obtained in this work

As tested on two important chemical steps (with very different FES) of the classic Strecker-cyanohydrin reaction for the synthesis of amino-acids in solution, the proposed protocol for optimal training set construction and NNP trajectory stabilization allows reproducing with excellent agreement the benchmark ab initio FES for a fraction of the computational effort.

We expect our approach to be easily generalizable to a range of chemical reactions in solution, allowing a limited and incremental use of costly AIMD calculations only if and when needed. In perspective, this controlled and efficient scheme will help to exploit NNPs to overcome a significant computational bottleneck in the accurate calculation of free-energy profiles in solution chemistry. In the next chapter, we will apply the methods presented in this chapter to another pathway of formation of glycine from formaldehyde, hydrogen cyanide and ammonia.

Chapter 4

Application to a NON-Strecker mechanism

4.1 Introduction

As explained before, glycine, the simplest amino-acid has a key role in prebiotic chemistry to understand the chemical evolution of life. Up until now, it has been observed in meteorites [142, 143, 144, 145, 146], but has never been observed in the interstellar medium (ISM) despite numerous attempts [147, 148, 149, 150]. This is why, the preferred proposed mechanism for its formation is the Strecker pathway presented in section 1.4. It is suggested in water ices in ISM, in water on primordial earth or at the heart of meteorites. Its main intermediate is aminoacetonitrile obtained by first the addition of ammonia on formaldehyde and then the addition of hydrogen cyanide. But there exists competitive mechanisms to the Strecker pathway.

First, the chemistry of the three components in water is very complex and has been extensively studied experimentally [151, 152, 18] and theoretically [153, 154, 155, 156]. In reference [153], the authors build the whole reaction network with 39 intermediates with the constraint that the intermediates must only have 2 carbon atoms. Their results are shown in figure 4.1, they used *ab initio* hybrid B3LYP simulations with implicit solvent. One striking fact in this figure, but also in experimental studies is that the formation of glycolonitrile (labelled 1B in figure 4.1) is considered as a mechanism opposed to the formation of aminoacetonitrile (labelled 38 on figure 4.1) and never as a precursor for glycine.

In this chapter, we propose a mechanism for the synthesis of glycine in water under prebiotic conditions passing by glycolonitrile instead of aminoacetonitrile. We use the method described in the previous chapter to sample the free energies along collective variables defined according to the protocol presented in section 1.4. Our results compare very well with the sparse literature available on the intermediates we found. Our findings explain the kinetic and thermodynamic behaviors observed experimentally.

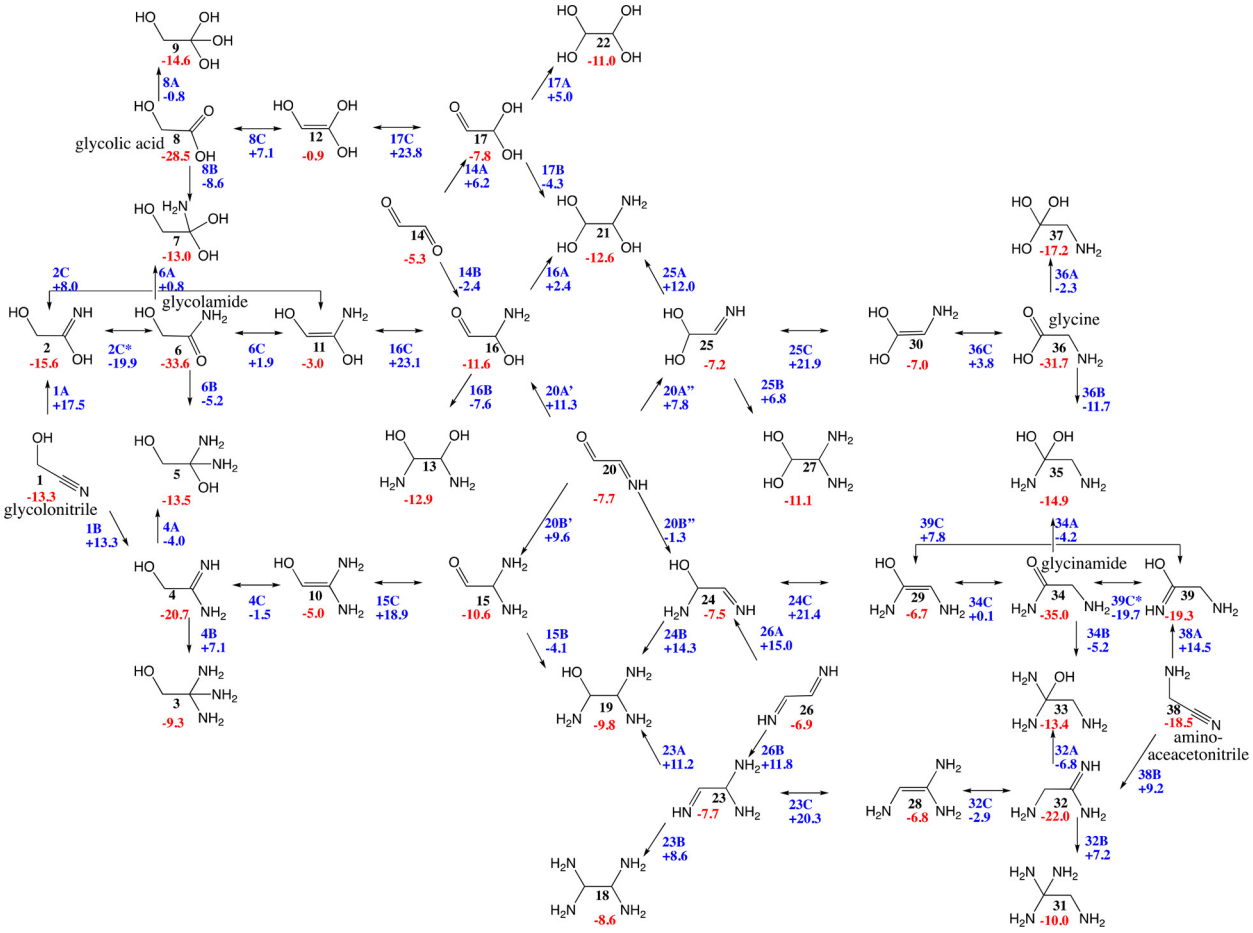

Figure 4.1: Reaction network of hydrogen cyanide, formaldehyde, and ammonia in water

4.2 Computational setup

To study this complex mechanism, we apply all the methods explained in the previous chapters. The metadynamics exploration steps and the building of the collective variables were carried out by Leon Huet (2nd year PhD student in the team), while I performed all the machine learning potential umbrella sampling simulations as well as the training of the potentials. The *ab initio* free energies were generated independently by Leon Huet to compare with the machine learning ones.

As collective variables, we used the (s, z) PCV with 12 reference frames obtained by shooting trajectories from putative transition states points as explained in subsection 1.4.2 for the most energetic steps of the mechanisms. For the steps only involving a proton transfer between an oxygen atom and a proton from the solvent, we chose to use only the hydrogen coordination number of the reactive atom. Indeed, in this case, only the coordination number of the oxygen atom of the reactive part would be important, and the coordination table would only have one line which would introduce more complexity for nothing. Moreover, using the coordination number instead of, for example, the distance between the reactive atom and a targeted atom of the solvent allows more flexibility for the reaction mechanism. These reactions were fully computed *ab initio* with no machine

learning, as only around 15 windows were needed to have a fully converged free energy profile.

In this chapter, to train machine learning potentials, we used the methods presented in chapter 3. As the starting reactants and the final product are the same, the simulation boxes are exactly the same, thus the free energy of the end state reported in this chapter should be the same as the one reported in ref [7].
Thus, the situation for this chapter is similar to the one of chapter 3, hence, the same

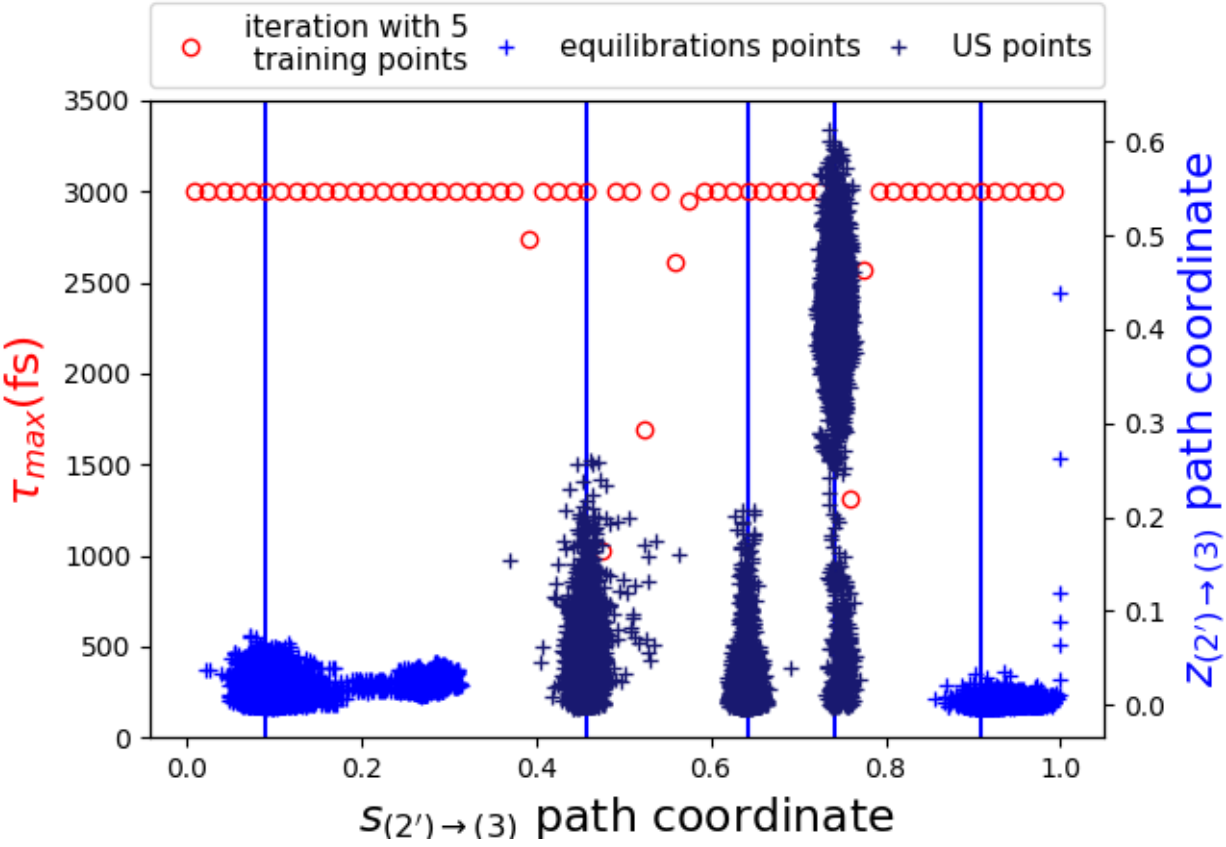

Figure 4.2: Performances of the MLP trained for step (2') → (3) of the mechanism. The points of the training set are represented

loss and the same protocol can be used to build a training set for each step of the mechanism. The main difference with chapter 3 is the fact that we did not have the results before performing the study, making it a study fully relying on the results of the MLPs for some steps of the mechanism. This also implies that the *ab initio* data can be subject to hysteresis effects that we chose to keep in the training set in order to have greater generalization performances. This is reported in figure 4.2, where the last training trajectory has a hysteresis effect on the z collective variable. This means that the system is exploring configurations that are far away from the pathway that we want to sample. This is a common problem in umbrella sampling simulations. In our case, it is most often

caused by protons interacting with water, or by the reaction occurring "too early" or "too late". Including these hysteresis effects in the training set will allow a bigger variety of configurations in the training set.

4.3 Results

4.3.1 The mechanism

Our proposed mechanism is shown in figure 4.3. Instead of the nucleophilic attack of the ammonia in the usual Strecker synthesis, here we study the nucleophilic attack of the hydrogen cyanide via the cyanide ion on the formaldehyde molecule to form, after a proton exchange with the solvent, the glycolonitrile. To bridge the gap between glycolonitrile and glycine, we found a very unstable compound in water (2-oxiranimine, intermediate (3)) that allows the nucleophilic addition of a water molecule (hydroxyacetamid intermediate (4)).

Following this, a nucleophilic substitution happens with a water molecule and the elimination of an ammonia molecule (glycolic acid intermediate 4') which will lead to the addition of an ammonia molecule towards the basic form of glycine (P). Due to the different intermediates, this mechanism is more favorable in basic medium. Indeed, the attack of the water molecule (HO- in basic medium) is on the less substituted carbon, while in acidic medium the attack would be on the more substituted carbon.

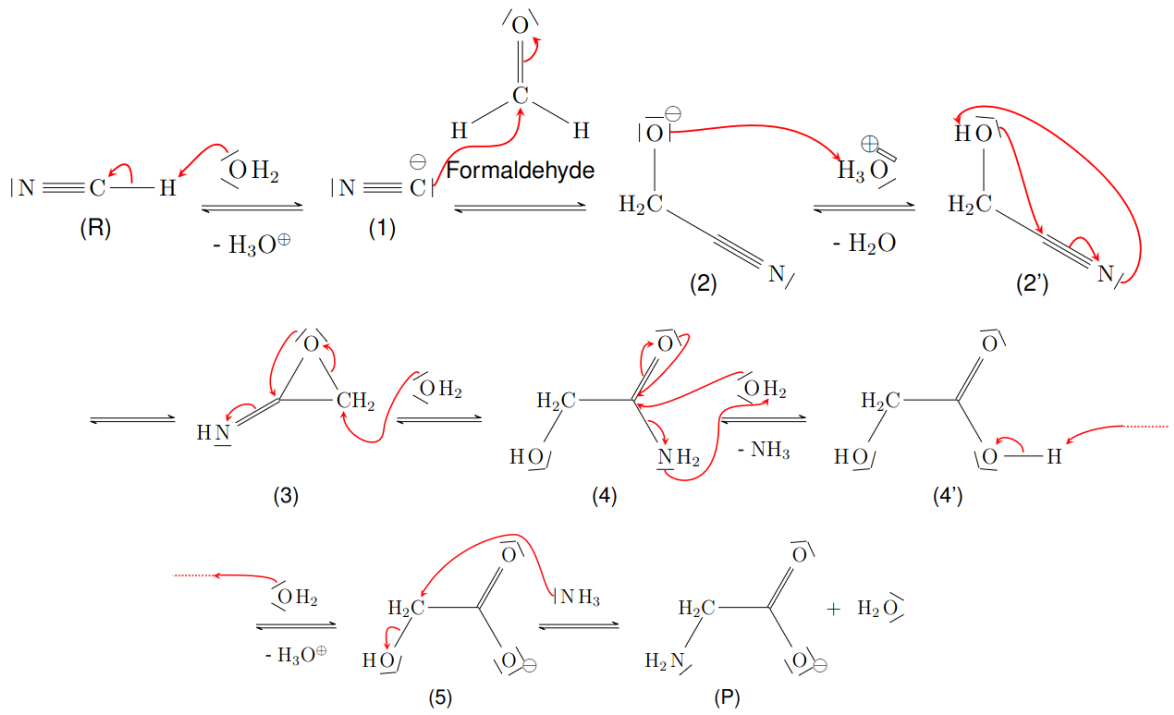

Figure 4.3: Our proposed mechanism of prebiotic synthesis of glycine

4.3.2 Free energies of the mechanism

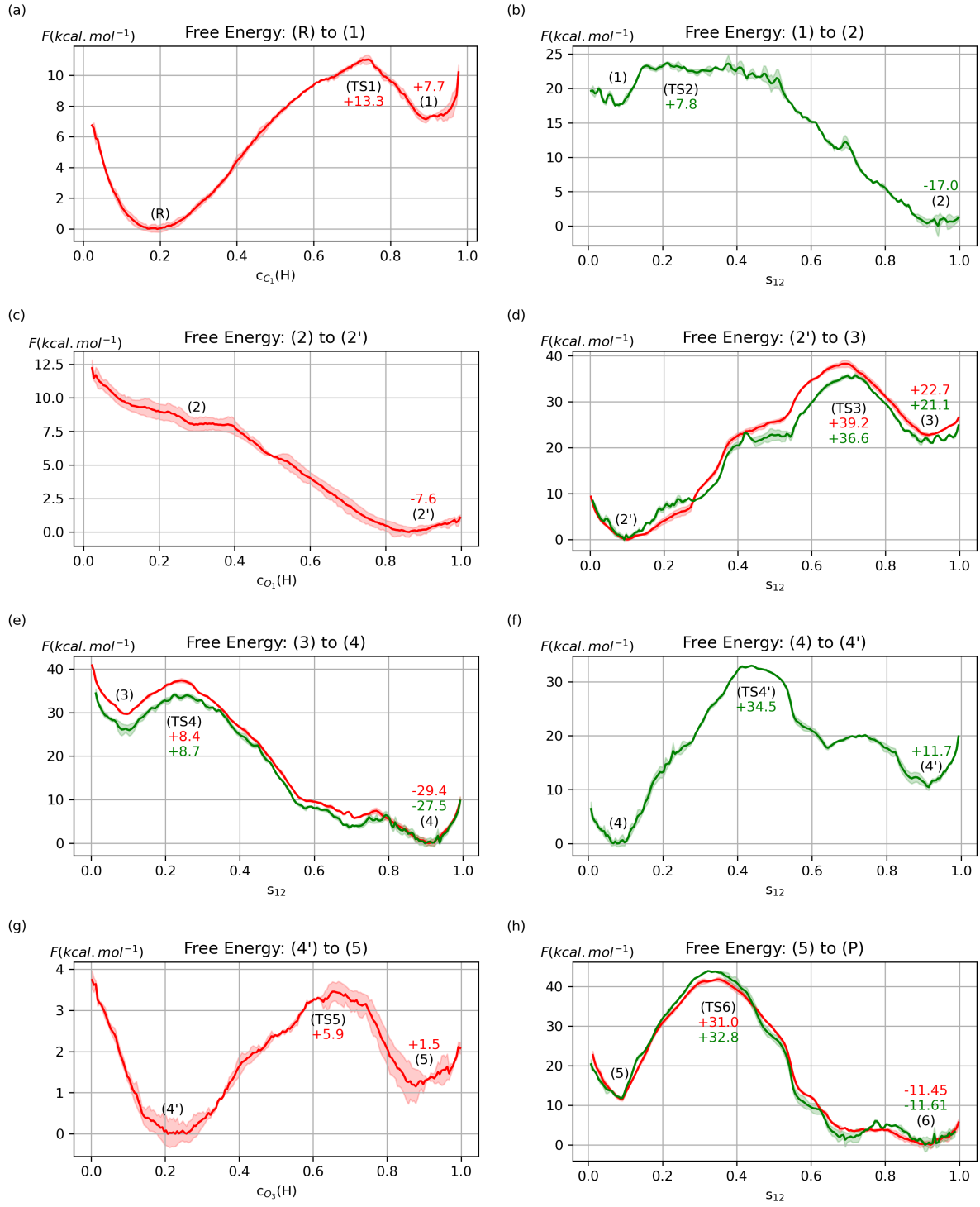

Figure 4.4: Free Energy profiles. In red the ones that were determined using only DFT, in green the ones that were determined using both DFT and neural network potential.

The free energy profiles of all the steps of the mechanisms are presented in figure 4.4. We chose as a convention to plot all the information related to *ab initio* calculations in

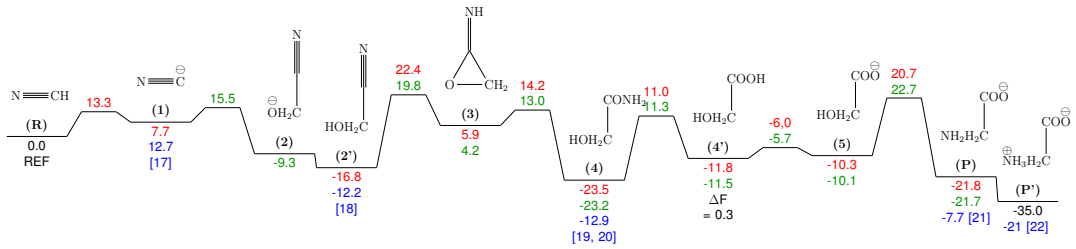

Figure 4.5: Balance of all the steps and all the reference free Energy encountered during the mechanism

red, while all the machine learning calculations are represented in green. As explained previously, since steps (R) → (1), (2) → (2') and (4') → (5) are just proton exchange between the reactive system and the solute, we chose to study those transformations only using AIMD and the hydrogen coordination number of the reactive atom as the collective variable.

The final balance of all the free energies is shown in figure 4.5. The values obtained with our protocol are compared with the experimental literature (in blue on the figure). The values of the proton transfers are obtained thanks to the pKa of the acid/base pairs. In the same fashion, we estimate the free energy difference between glycolic acid (4') and glycolate (5) that is 5.2kcal/mol [17]. This value is higher than the one found in the literature due to the electronic structure calculations. But, in most of the cases where data is available, our results compare very well with the experimental ones. Finally, some other theoretical studies have investigated the formation of glycine from formaldehyde and hydrogen cyanide. Most of the literature only considers α aminoacetonitrile as a precursor of glycine and the formation of glycolonitrile as a competing product.

Our results are also close to the ones presented in reference [153] and reported in figure 4.1, the difference in free energy between our method and theirs could be explained by the difference of methodology, indeed, they use implicit solvent and perform static computations, while we have explicit solvent, detailed mechanism and dynamic behavior.

4.3.3 Prebiotic relevance of the mechanism

It is very hard to find glycolonitrile in the ISM [157], but products from its photodecomposition have been observed [158], indicating that this reaction may occur in the ISM. Carbonaceous chondrites formed in the protosolar clouds from silicates and iron are in close interaction with the ISM, ices, and dusts in the forming solar system. The accretion allows the formation of chondrites with water, leading to water to rocks ratio from 0.2 to 0.7 [159, 160, 161] with organic matter. Thus, glycolonitrile may be formed in the ISM and then incorporated in the formation of chondrites and the rest of the mechanism

may occur within the chondrite as a big amount of solution chemistry happens in parent bodies of chondrites [162, 163, 164].

From this work and the work of Théo Magrino on the Strecker synthesis [7] we may conclude that glycolonitrile is less stable in water than aminoacetonitrile, but forms more rapidly. This is also the conclusion from an experimental study of the competing mechanism: the more ammonia in the solution, the more stable aminoacetonitrile will form [151]. This might explain why glycolonitrile cannot be found in meteorites as it is the kinetic product and at the timescale of the formation of the solar system, the thermodynamic product (aminoacetonitrile) will form, while glycolonitrile might react further on.

The intermediate oxiranimine is very unstable and there is no chance of observing it in meteorites or in water solution, but it leads to glycolic acid that is more stable than glycolonitrile and that has been observed in meteorites [165, 166, 167]. Therefore, this mechanism would explain the extraterrestrial delivery of organic materials such as glycolic acid or glycine on earth. Another lead for the appearance, development, and evolution of organic matter on earth is the idea of Charles Darwin followed by Oparin and Haldane of a primordial soup with reactions happening in a "one pot" fashion. Although we do not have the information regarding the concentration of ammonia necessary for this new mechanism, this mechanism may be a good alternative to the Strecker pathway where a higher concentration of ammonia is needed. Finally, glycine could also originate from electric discharge in an atmosphere composed of simple molecules, here again, the Strecker pathway is a hypothesis, but some work show alternative pathway [54], our mechanism could also be an alternative.

4.4 Conclusion

In this part, we have set up a method to train an accurate machine learning potential to study chemical reactions in solution. This method is based on a committee method: by training several neural networks, the deviation on the prediction of the forces can be tracked, and a simulation lifetime can be defined by determining the time at which the deviation on the forces exceeds a certain threshold. This simulation lifetime allows tracking the zones of the RC-space in which the machine learning potential is not well-trained and to target new training points. We use this method to sample the free energies of an alternative mechanism to Strecker synthesis of glycine. Our findings are in close agreement with the experimental data, but also allow explaining several observations of the meteorites.

Part III

Towards an agnostic description of chemical mechanisms

Chapter 5

Theoretical interlude: towards a better understanding of chemical mechanisms

In the previous chapters, we introduced a way to train machine learning potentials to perform umbrella sampling simulations. The training set was built using *ab initio* constrained molecular dynamics around a predefined point of the RC-space. This process thus requires the prior knowledge of the transition mechanism and the definition of a RC. Therefore, if the RC is bad and does not describe the transition mechanism well, the machine learning potential will be bad, but also the sampling of the free energy. Thus, a new RC needs to be built, and with it a new machine learning potential which increases the computational cost. These considerations raise several questions: what is a good collective variable? How do we assess its quality? How can we explore the transition mechanism in an unbiased way? In the following chapter, we introduce different theoretical tools to answer these questions. Then, in the second section, we introduce a theoretical way to compute kinetic rates.

5.1 Collective variables: understanding the transition mechanism

As explained before in this manuscript, a collective variable is the projection of the 3N atomic coordinate onto a variable. This projection should be done in order to grasp the whole complexity of the transition mechanism. Let us see what happens for a very simple system.

The following analysis is from [168]. In this work, the authors study the free energy profile of a double well potential along different coordinates. Here, the analytic free energy profile is known, and it is a perfect toy model to illustrate the dramatic effect of a wrong projection on the estimation of the kinetic rate and the free energy barrier. The results obtained in ref [168] are reported in figure 5.1. First, they project the dynamics on the "best" collective variable where the analytical barrier is, and they manage to recover the right free energy barrier of the reaction. Then, they project the dynamics on a linear

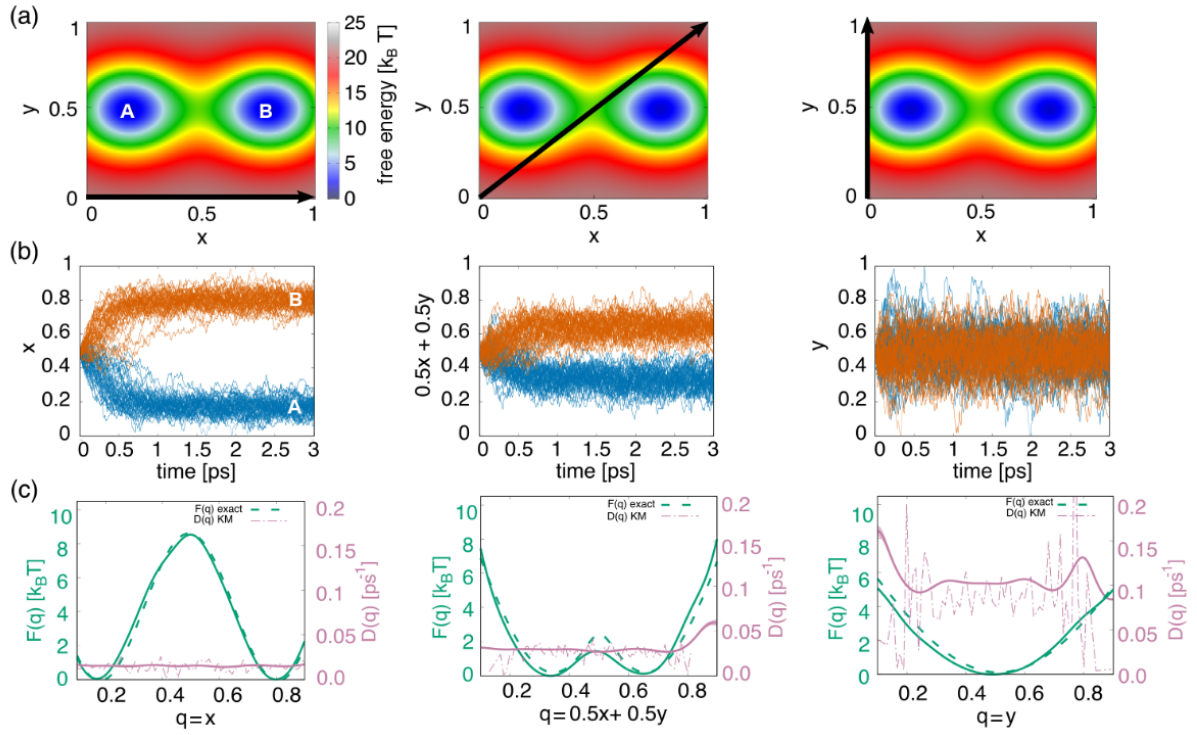

Figure 5.1: a) Analytic double-well free-energy landscape, with arrows indicating one-dimensional projections along the x-axis, the diagonal, and the y-axis, respectively. (b) 100 Langevin bi-dimensional trajectories, projected on a single CV, relax from the barrier-top into A (blue) or B (orange), and are used to optimize effective one-dimensional Langevin models via likelihood maximization. (c) For each projection, the free-energy and diffusion profiles inferred from the latter models are shown (continuous lines with standard error over 10 reconstructions) alongside the exact F (q) profiles and an alternative estimate of D(q) (dashed lines). Results adapted from [168]

combination of the two coordinates (panel b of figure 5.1), a barrier and the two metastable states can be identified, but the barrier is underestimated compared to the previous case, meaning that some information is lost in the projection process. Finally, the authors project the dynamics on the y coordinate, where no transition happens. On figure 5.1, panel c), there is only one state and no identified barrier. Therefore, all the information about the transition is lost and y can be qualified as a "bad collective variable".

In terms of chemical reactions in solution, there exists reactions where the solvent plays an important role compared to the gas phase behavior. For example, the solvent might create some steric interactions that can hinder the formation of the transition state, leading to a higher free energy barrier. But, these solvent degrees of freedom need to be taken into account in the collective variable, therefore, a general, automatic way of generating collective variables can be used.

For simple reactions implying only a proton exchange between the reactive system and the solvent, only the hydrogen coordination number of the reactive atom can be used, as it

was done in the step (0) → (1) of the previous chapter. But, for more complex reactions, where several atoms are involved, or reactions where a solvent atom is not directly used, the coordination is not enough to include the solvent degrees of freedom. A first solution, is the one taken in this thesis, where, the coordination of the reactive atoms with respect to all the species in the system are taken to build a path collective variable. Another new way of building high quality collective variable including exterior effects is by using machine learning methods. [169, 170, 171].

Now that we have seen what are good collective variables in terms of projection, we will introduce the definition of the committor probability, how it is useful and how it allows assessing the quality of a collective variable and to devise collective variables.

5.1.1 The committor probability

In a system with N atoms, two metastable states A and B separated by a free energy barrier, we can compute the probability $\phi_B(\mathbf{x})$ of a given atomic configuration $\mathbf{x}$ to end up in state B before state A given it started in $\mathbf{x}$. This probability is called the committor probability, and was first introduced by Onsager in 1938 [172]. Numerically, it can be approximated by starting m configurations at $\mathbf{x}$ with random velocities drawn from the Boltzmann distribution and determine in what state the simulation ends. Let n_B be the number of simulations ending in B, then, the committor probability is approximated as:

$$\phi_B(\mathbf{x}) \approx \frac{n_B}{m} \tag{5.1}$$

Because high free energy barriers mean low probability of being out of the metastable states, assessing ϕ_B for rare events needs extra precaution, indeed many configurations will have a committor value close to zero or one, thus to assess it accurately, many trajectories are needed. But, this quantity gives perfect information and exactly what is expected from a collective variable: the commitment probability can be seen as a measure of the progress of the reaction. Moreover, this allows to precisely define what a transition state is: a configuration is said to be a transition state if a trajectory started from this configuration has the same probability to end up in both metastable states.

In other words, a transition state is a configuration that has a committor value of 0.5 [173, 174]. Even though, the committor is said to be the best collective variable, it has no physical interpretation and does not give insight about the transition mechanism [175], therefore, it can be used as a tool to analyze data. Moreover, it is impossible to use it to bias simulations.

5.1.2 Committor analysis

From the definition of the committor, it follows that for an optimal collective variable, a single value of the committor should be mapped to a single value of the collective variable [176]. This is in practice never the case, but it can come close to this.

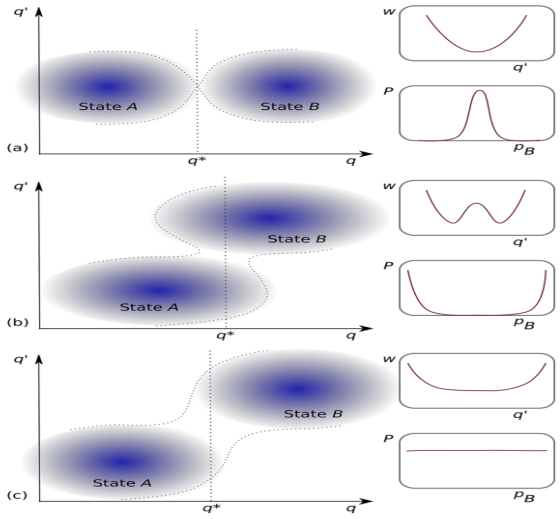

Figure 5.2: Illustration of the committor analysis procedure to assess the collective variable quality along different free energy landscapes. Figure adapted from [175]

To assess the quality of a collective variable, one can perform a so called "committor analysis". To do so, configurations are gathered around a value of the CV, q^* that is the putative value of the CV for the transition state. The value of the committor is estimated for these configurations and the histogram $P(\phi_B(q^*))$ is computed. The summary of the situation can be found in figure 5.2 that is adapted from ref [175], where three different double well free energy surfaces are explored along the q coordinate, for the first one, all the information is on the q coordinate, and therefore, the sampling of the transition mechanism is perfect, and the committor analysis yields to a distribution sharply peaked around 0.5. In the second case, the free energy has a component along q', therefore, the sampling is not optimal and yields to a committor distribution peaked around 0 and 1, because, the line q^* crosses the basins and almost does not pass by the transition region. Finally, in the last case, the barrier is flat in the q' direction which leads to a flat distribution of the committor.

In this thesis, committor analysis was used to generate the CV, in a very pedestrian way. As explained in subsection 1.4.2, unbiased trajectories are shot from the putative pathways and reference frames are chosen on these trajectories. However, only a few runs are performed due to the computational burden of *ab initio* calculations, and this makes it a heavy procedure to identify only a few transition states. A better way to identify transition states and transition paths are transition path sampling (TPS) algorithms.

We also used committor analysis to assess the quality of a collective variable in chapter 6 with the help of a machine learning potential

5.1.3 Transition path sampling

Transition path sampling algorithms [23, 177, 178, 179, 180] are a set of unbiased methods used to explore transition path. They are stochastic Metropolis algorithms for which

the sampling is made in trajectory space instead of the atomic position space for other enhanced sampling techniques presented in this thesis. Moreover, they are unbiased techniques, meaning that the dynamics and the mechanism of the transition can easily be accessed. Here, we will present the shooting from top algorithm [179] that was used in this thesis that is a variation of transition path sampling. First, start from a first raw transition trajectory obtained by any means. Oftentimes, it is obtained with metadynamics

- Define the "shooting range" $S = [s_A, s_B]$ from which shooting points are selected
- Randomly choose a shooting point in S with uniform probability in the last accepted trajectory
- Draw velocities from the Maxwell Boltzmann distribution (at 300 K) and propagate two trajectories, one forward in time and one backward, until they reach reactants or products.
- Accept the shooting if one trajectory reached the reactants and the other the products with probability $Min(\frac{n}{n'})$ where n and n' are respectively the number of points of the previous (resp current) trajectory.
- iterate

A schematic summary of this algorithm is presented in figure 5.3, adapted from the thesis of Alexandre Jedrecy [181] in which he used transition path sampling algorithms to investigate the liquid-liquid transition in water's no man's land. In this thesis, we use a variation of the original TPS algorithm by using the "shooting from top" [179] scheme, where the configurations are exclusively selected in the $S = [s_A, s_B]$ range. For this method, we are only interested in Hamiltonian trajectories.

This method allows having an unbiased exploration of the transition mechanism and to gather transition pathways. Moreover, this method allows an exploration of the chemical space without relying on a collective variable and hence does not need a prior knowledge of the reaction mechanism. Indeed, even though a coarse variable is needed to locate the top of the barrier, the range of values was shown not to be important. This shooting from top scheme is used to be more efficient than traditional transition path sampling techniques like aimless shooting, where the new configuration was chosen in the whole previous trajectory. Although it is a new technique and has not extensively been used, from the experience in the group, it seems to be a good method to save computational time compared to traditional TPS.

5.2 Kinetics

Until now, we have mainly presented tools to recover thermodynamic properties of chemical mechanisms. But, in such events, there are always two competing concepts: thermodynamics that will tend to push the system towards the most stable state and kinetics that will favor the pathway that forms the products the more rapidly. In other words, in

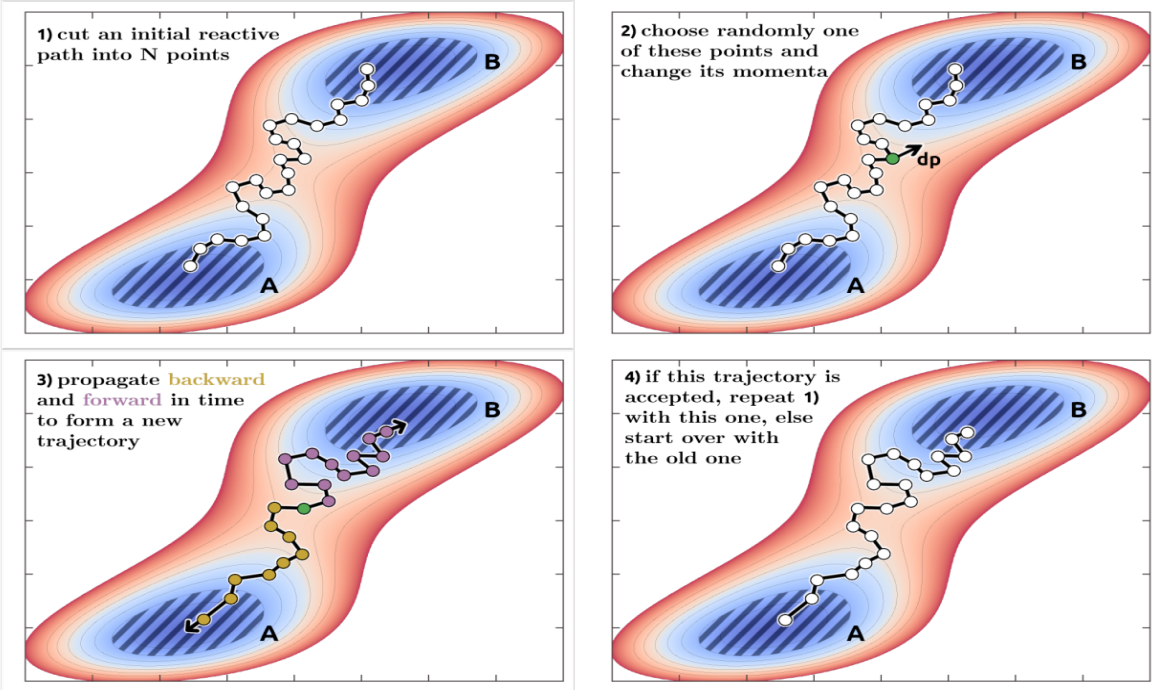

Figure 5.3: Schematic illustration of the transition path sampling algorithm, adapted from [181]

theory, from a macroscopic point of view, if one waits a small amount of time, the kinetic product (the one with the smallest activation barrier) will form, while if one waits for an infinite amount of time, the thermodynamic product (the most stable one) will form. In this section, we will present the macroscopic description of the kinetic behavior of a reaction, then, we will show the microscopic behavior and make a link between the two.

5.2.1 Macroscopic point of view

Consider a unimolecular reaction with two states A, and B separated by a barrier. We want to know the speed at which A transforms in B and B transforms in A, in other words, we want to know the kinetics of the reaction. The density in A/B is noted $C_{A/B}$, the phenomenological equations describing the combined evolution in time of C_A and C_B are:

$$\begin{cases} \dfrac{dC_A(t)}{dt} = -k_{A\to B}C_A(t) + k_{B\to A}C_B(t) \\ \dfrac{dC_B(t)}{dt} = -k_{B\to A}C_B(t) + k_{A\to B}C_A(t) \end{cases} \tag{5.2}$$

Where, here, we make the assumption that the rate of transformation is proportional to the concentration of the state. $k_{A\to B}$ and $k_{B\to A}$ are the two rate constants.

From this set of equations, it follows that the total density is conserved, *i.e.*, by summing the two equations:

$$\frac{d(C_A(t) + C_B(t))}{dt} = 0 \tag{5.3}$$

This comes with the assumption that the overall quantity in the system is conserved and there is no external intervention. At equilibrium, when the variation of the concentration is zero, we find using the equations:

$$K = \frac{\langle C_A \rangle}{\langle C_B \rangle} = \frac{k_{B\to A}}{k_{A\to B}} \tag{5.4}$$

K is the equilibrium constant, characterizing the relative stability of the states. the angular brackets indicate the equilibrium densities of each species. In general, in the lab, the system is prepared in one initial state, containing only one species for example:

$$\begin{cases} C_A(0) = C_A \\ C_B(0) = 0 \end{cases} \tag{5.5}$$

This system of equation can easily be solved:

$$\begin{cases} C_A(t) = \langle C_A \rangle + \left[C_A(0) - \langle C_A \rangle\right] e^{-(k_{A\to B}+k_{B\to A})t} \\ C_B(t) = C_B \left(1 - e^{(k_{A\to B}+k_{B\to A})t}\right) \end{cases} \tag{5.6}$$

These equations allow us to define a relaxation time:

$$\tau_{rxn} = (k_{A\to B} + k_{B\to A})^{-1} \tag{5.7}$$

which, can be written in terms of density using equation 5.4:

$$\tau_{rxn}^{-1} = k_{A\to B} \frac{\langle C_A \rangle + \langle C_B \rangle}{\langle C_B \rangle} \tag{5.8}$$

In an ideal case, one can obtain $k_{A\to B}$ by counting the number, n_{AB} of transitions between A and B during a "long enough simulations", then, $k_{A\to B}$ can be assessed:

$$k_{A\to B} = \frac{n_{AB}}{t_A} \tag{5.9}$$

Where t_A is the total time spent in state A during the simulation. But as explained in section 1.3.4, in the case of chemical reactions, the free energy barrier is too high and not enough transitions are observed to accurately compute the transition rate.

To overcome this problem, and to make a link between the macroscopic phenomenological behavior described above, and the microscopic description offered by molecular dynamics simulations, some statistical tools have been put into place.

5.2.2 Transition state theory

TST [25, 90, 182, 183] is a mean of getting an approximate kinetic information easily from thermodynamics. At the center of the theory is the concept of transition states. To go further into transition state theory, we must first define a reaction coordinate q that will track the evolution of the reaction. We define as transition states the points

on the orthogonal dividing surface crossing the top of the barrier. The transition rate can be computed by taking the ratio of the population at the transition state Z^* and the population at the initial state Z_A multiplied by the velocity at which the trajectories cross the separating surface. But, this method does not take into account the trajectories that cross the dividing surface, but then come back to the initial state. To overcome this problem, the "recrossing factor" κ is introduced. It is set to one in the frame of TST, in the end, the transition rate is given by:

$$k_{A\to B} = \kappa\langle v^*\rangle \frac{Z^*}{Z_A} \tag{5.10}$$

Where $\langle v^*\rangle$ is the mean velocity of the collective variable at the dividing surface. Every quantity, except for, κ can be computed using umbrella sampling simulations, This is why TST is most often used to compute kinetic information, since the leading term is Z^* that depends on the exponential of the activation barrier. Since TST does not take into account the recrossing events, κ is always less than one and TST always overestimates the transition rate.

5.2.3 Reactive flux formalism

To accurately compute the transmission coefficient κ, Bennett and Chandler [25] have come up with statistical tools based on trajectories started near the top of the barrier. Using this, we have access to the dynamical behavior of the system. We first define a collective variable q that will track the evolution of the reaction. It will discriminate whether the system is on the reactants side of the barrier or on the products side. To do so, we define a transition state q^* and the two following functions:

$$\begin{cases} h_A(t) = \theta(q^* - q(t)) \\ h_B(t) = \theta(q(t) - q^*) \end{cases} \tag{5.11}$$

These two functions are the indicator functions, they describe the state of the system. Using them, we define the time correlation function which is the proportion of trajectories in state B at time t given it started in state A:

$$\mathcal{C}(t) = \frac{\langle h_A(q(0))h_B(q(t))\rangle}{\langle h_A\rangle} \tag{5.12}$$

$\langle h_A\rangle$ can be linked to the equilibrium densities of the previous section:

$$\langle h_A\rangle = \frac{\langle C_A\rangle}{\langle C_A\rangle + \langle C_B\rangle} \tag{5.13}$$

From all of the above, is the system is initially prepared in state A, the time evolution of the products density is:

$$C_B(t) = (\langle C_A\rangle + \langle C_B\rangle)\mathcal{C}(t) \tag{5.14}$$

If the time is larger than the typical molecular timescale (τ_{mol}) equations 5.6 and 5.14 can be combined to find:

$$\mathcal{C}(t) = \langle h_B\rangle(1 - e^{-(k_{A\to B}+k_{B\to A})t}) \tag{5.15}$$

Now, we are interested in events from time before τ_{rxn} therefore, we can write

$$\mathcal{C}(t) \approx \langle h_B \rangle (k_{A\to B} + k_{B\to A})t \tag{5.16}$$

Which, can be written using equations 5.8 and 5.13:

$$\mathcal{C}(t) = k_{A\to B}t \tag{5.17}$$

Hence the rate constant for $\tau_{mol} < t < \tau_{rxn}$:

$$k_{A\to B} = \frac{d\mathcal{C}(t)}{dt} \tag{5.18}$$

We now have a way to link macroscopic behavior of densities and phenomenological rate constants with microscopic quantities that can be computed using molecular dynamics simulations. This result can also be found by using the fluctuation dissipation theorem in the framework of linear response theory [184]. Now, we will show how to compute this quantity using molecular dynamics simulations in the Bennett-Chandler formalism. By using time translation invariance and the definition of $\mathcal{C}(t)$ one gets:

$$k_{A\to B} = \langle \dot{q}(0) h_B(t) \rangle_{q(0)=q^*} \frac{\langle \delta(q(0) - q^*) \rangle}{\langle h_A \rangle} \tag{5.19}$$

The second term of this equation can be easily computed using Umbrella Sampling simulations, while the second term is the one that characterizes the deviation from TST, indeed, it is clear that here the first term takes into account the whole behavior of the system and not just the first crossing of the separatrix by the system. However, this term needs a large amount of trajectories to reach statistical convergence, this is why TST is often preferred, especially in *ab initio* studies where most of the computational effort is put on the sampling of the free energy.

Chapter 6

Agnostic machine learning description of chemical reaction in solution

In the previous chapters, we introduced theoretical tools to study chemical reactions in solution from a thermodynamics point of view, but also from a kinetic point of view. We also devised and applied a method to study chemical reactions in solution with the help of machine learning, but this method needed the prior definition of a collective variable. This means that the transition mechanism has to be well understood, which is sometimes a tough task with *ab initio* simulations. In this chapter, we introduce a way based on transition path sampling to train a machine learning potential with which it is not only possible to recover the free energy profile that gives the thermodynamical information but also the kinetic rates of a reaction. We apply this method to the prototypical benchmark system of the nucleophile substitution of methyl chloride with a chlorine ion.

6.1 Introduction

In this chapter, we once again rely on the data generated during Theo Magrino's thesis [24] to build a machine learning potential and compare its results with the *ab initio* standards. We go a bit further by computing the kinetic rates and performing some committor analysis. To do so, during his thesis, Theo Magrino used transition path sampling and enhanced sampling techniques to study a prototypical reaction. This reaction is the S_N2 substitution of the methyl-chloride molecule with a chlorine ion, in other words, nothing changes because the chlorine atom of the molecule is replaced by another chlorine. But, this reaction has widely been used to study the impact of the solvent [66], the impact of the DFT functional [185], and the effect of the collective variable taking or not into account the solvent degrees of freedom [186].

Indeed, in gas phase, the reaction has two wells that disappear in the solvent case because the ion/molecule complex is stabilized in gas phase due to long distance interactions, but this disappears due to the cost of desolvatation of the components [66, 187, 188, 189]. Furthermore, in gas phase, the transition state is stabilized due to the negative charge

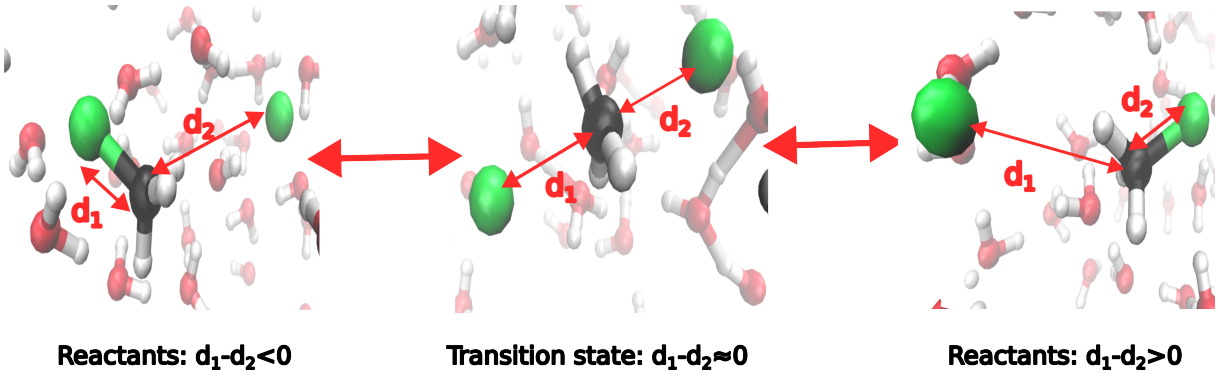

Figure 6.1: Illustration of the methyl chloride substitution reaction along with the typical collective variable used: $d_1 - d_2$.

that is split between the two chlorine atoms, which is the contrary in the reactants or products well since the charge is carried by only one ion. Adding a solvent will destabilize the transition state and increase the activation barrier [190, 191].

The reaction is presented in figure 6.1, along with the two most important quantities of the reaction, d_1 and d_2 that are the distances between the chlorine atoms and the carbon atom. The most simple collective variable one can think of is $d_1 - d_2$, in the reactants part, it will be less than zero (the carbon atom is linked to one chlorine atom), around the transition state its value will be close to zero, and on the products side, it will be greater than zero (the carbon atom is linked to the other chlorine atom). Although it was shown that for some system like the ion pair dissociation the solvent degrees of freedom needed to be taken into account, for this particular reaction, every study indicates that $d_1 - d_2$ is one of the best CV possible.

In this chapter, we will first present the training set and how it was generated, then, we will present how our model behaves when sampling the free energy profile along two different collective variables. We then present our kinetic assessment of the transition mechanism along the two same collective variables. Finally, we will judge the quality of the training set by performing committor analysis.

6.2 The training set

In ref [24], Theo Magrino and Léon Huet performed transition path sampling simulations on the reaction described above. First, the reactants and products are defined, by performing long *ab initio* equilibrations runs in the two basins. From these runs, the coordination numbers of the reactive atoms: the carbon and the two chlorine are averaged over the two runs to build two coordination tables The chosen coordination tables are shown in figure 6.2.

After this, these two tables are used to define a coarse path collective variable s_2. s_2 is built to approximately locate the top of the barrier in order to apply the shooting

		Set of atoms σ				
		C	Cl	O	K	H
Atom i	C	0.0	0.85	0.18	0.03	2.91
	Cl_1	0.85	0.00	0.58	0.01	1.01
	Cl_2	0.0	0.00	0.90	0.00	1.64

(a) Ref table for reactants $CH_3Cl_1 + Cl_2^-$

		Set of atoms σ				
		C	Cl	O	K	H
Atom i	C	0.0	0.85	0.18	0.03	2.91
	Cl_1	0.0	0.00	0.90	0.00	1.64
	Cl_2	0.85	0.00	0.58	0.01	1.01

(b) Ref table for products $CH_3Cl_2 + Cl_1^-$

Figure 6.2: Reference coordination matrices of two-state s_2 and z_2 path CVs employed on metadynamics simulations. Values are averages over reactants equilibration, and chlorine lines are switched to represent products. Coordination numbers are given for a given atom i with respect to a set σ, corresponding to all atoms of the same element.

from top scheme presented in subsection 5.1.3. It was however showed that the efficiency of the algorithm did not depend on the chosen range for the transition state region. In this study, we chose to shoot trajectories in the interval $[1.35, 1.65]$. In this way, a large amount of unbiased trajectories bridging the reactants and the products basins can be obtained easily. This gives us a variety of transition paths and transition states that describe well the transition path ensemble. In figure 6.3 the TPS trajectories are projected on two different CV: $d_1 - d_2$, and s_2 that is used to define the shooting range.

These trajectories contain different unbiased reactive paths that should grasp the whole

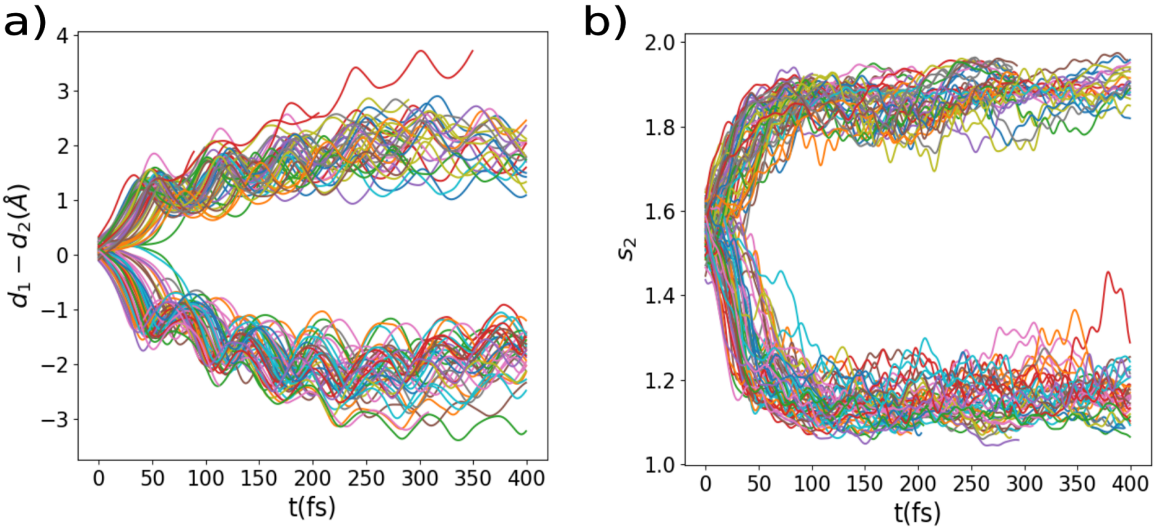

Figure 6.3: TPS training trajectories. Panel **a)** TPS training trajectories as a function of time and the heuristic collective variable: the difference between the two clorine-carbon distances, $d_1 - d_2$ along with panel **b)** the same trajectories as a function of the simple path collective variable s_2

diversity of the atomic environment of the reactive atoms and the solvent during the reaction. This is the key to training a MLP which relies on the representation of local atomic environments. Therefore, we used these TPS trajectories as training data along with the method presented in chapter 3. To do so, we use the deepmd smooth-edition package [10, 120] which is based on a Behler-Parrinello [9] structure.

To deal with the heterogeneity of the system that is intrinsic to reactions in solution and to avoid water molecules to have an overwhelming weight in the training of the potential, we use a custom loss function to optimize the neural networks weights given by equation 3.1.

6.3 Assessment of the training set

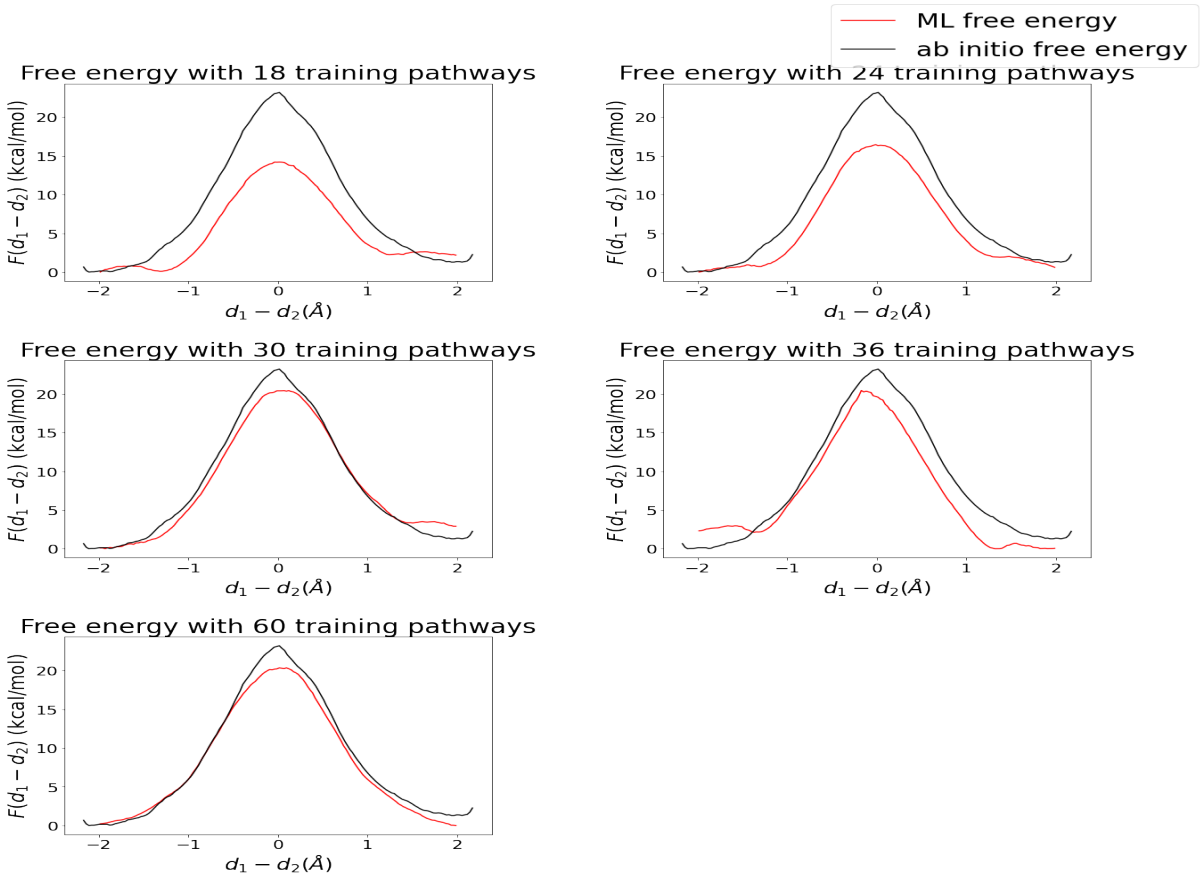

Figure 6.4: Free energy obtained along the $d_1 - d_2$ collective variable with increasing size of the training set (red) compared with the *ab initio* reference

Finally, we assess the quality of the training process by training neural network potentials with an increasing quantity of transition pathways. The results are shown on figure 6.4, it is now clear that if the number of training trajectories in the training set is too low, the sampling will be insufficient leading to an underestimation of the barrier height as can be seen with the plots with 18 and 24 training pathways. Then, the free energy profile reproduces quite well the *ab initio*, but we decided to keep increasing the number of training points for two reasons: first, as seen in 6.4, the obtained free energy profile matches quite well the *ab initio* one except for one asymmetry between the level of the reactants and the one of the products which is problematic for a symmetric reaction. Moreover, we noticed that the training could be improved by looking at the stability of the NNP: for some windows, too many "mirror reflections" were observed, meaning that

the system was stuck in some kind of "uncertainty bubble" in which it could not get out. In the end, we chose the neural network potential with 60 training pathways, and it is the one we use in the following sections of this chapter.

6.4 Umbrella sampling

6.4.1 Free energy along $d_1 - d_2$

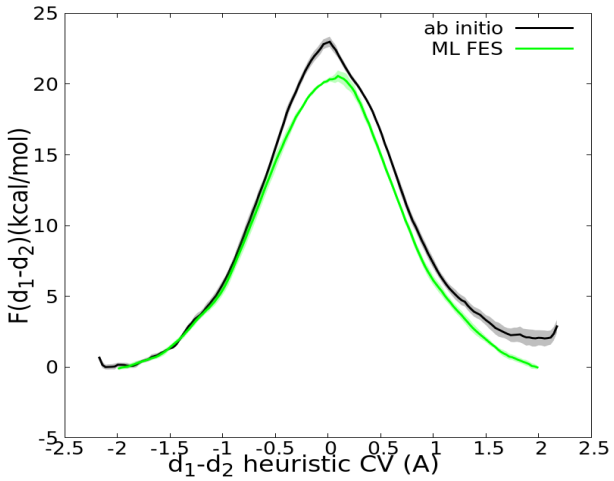

Figure 6.5: Free energy along a heuristic collective variable based on distances computed using *ab initio* umbrella sampling simulations (black) and computed using machine learning potential based umbrella sampling simulations (green). The shaded zones correspond to the estimated statistical errors.

As it is the most commonly used collective variable in this reaction, we started our study by performing umbrella sampling simulations along $d_1 - d_2$ to compare it with the *ab initio* data we already had, but also with the literature. We equally split $d_1 - d_2$ in 60 equally spaced windows with a spring constant of 1.14288 eV between $d_1 - d_2 = -2$ Åand $d_1 - d_2 = 2$ Å.

In order for the neural network potential to be stable, a semi-parabolic constraint was added on the chlorine-chlorine coordination number because it was observed that during the simulation, the two atoms were getting too close. This behavior is completely non-physical, as the two atoms are negatively charged. This is most likely due to the fact that configurations such that the two chlorine atoms are too close to each other are not in the training set, and including these high energetic frames in the training set would risk biasing the training towards these chemically uninteresting configurations. This is why we chose to proceed this way. It was implemented similarly in a MLP with a repulsive term to avoid the atoms to collide [192].

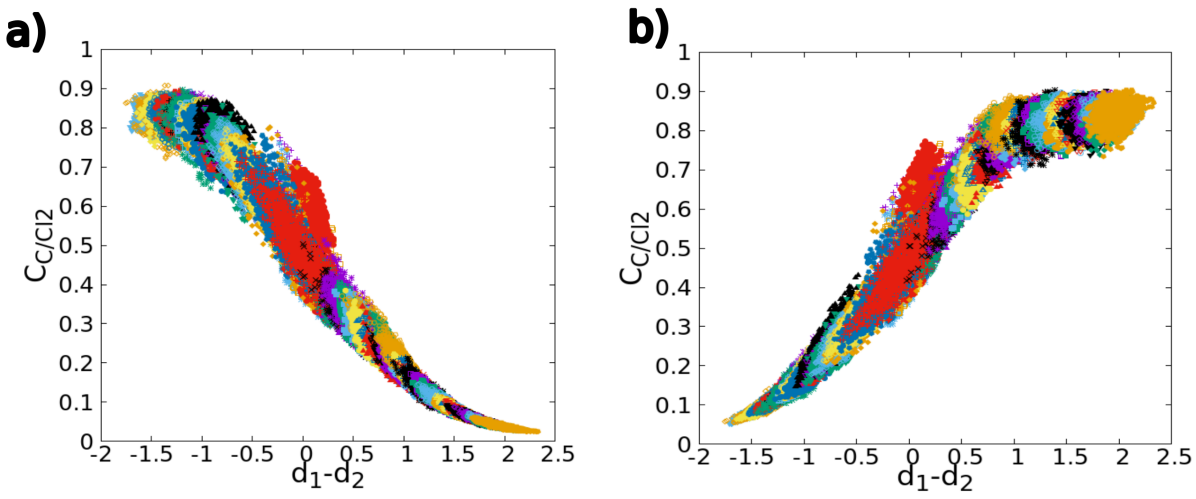

Figure 6.6: Umbrella sampling simulations trajectories projected on the C/Cl1 and C/Cl2 coordination numbers

The results are presented on figure 6.5, in the *ab initio* case and in the machine learning case simulations were run for 20ps. The simulations were cut in four, the two first fourth were discarded, and the two last were kept. The mean of the two lasts was taken as the free energy, while the difference was considered as the statistical uncertainty. The agreement between the *ab initio* simulations and the machine learning is satisfactory, given that the training was performed with only transition path sampling simulations and no umbrella sampling data. The simulation points are represented in figure 6.6, no holes are observed in the sampling, and no hysteresis effect either, meaning that the sampling is satisfying. This shows that this method is more generic than the previous one and more powerful.

6.4.2 Path collective variable

Since our training method does not depend on a collective variable, we can try to sample the free energy of the reaction along a different CV. In ref [24] the authors use transition path sampling to find reference frames for building PCV. But, with *ab initio* molecular dynamics, the computational burden can be very heavy, since after performing transition path sampling simulations, one also has to perform the expensive umbrella sampling simulations.

This is why, we performed many TPS simulations with our MLP, and chose reference frames among around 5000 TPS simulations. These frames were chosen according to a minimizing metropolis Monte-Carlo method based on a nudged elastic band (NEB) technique presented in ref [24]. The aim is to have reference frames that are equally spaced in the PCV metric space. First, N=10 frames are chosen among the TPS configurations, then a fictitious NEB energy is defined in equation 6.1 and minimized using a Monte-Carlo

procedure by choosing a new reference frame between k and k+1 at each iteration.

$$E = \sum_{k=1}^{\mathrm{N}-1} \left(\mathrm{D}_{k,k+1} - \frac{1}{\mathrm{N}-1} \sum_{l=1}^{\mathrm{N}-1} \mathrm{D}_{l,l+1} \right)^2 + \beta \sum_{k=2}^{N-1} [\max(\theta_k - \theta_{tresh}, 0)]^2 \tag{6.1}$$

Where, $D_{k,k+1}$ is the distance between frame k and k+1 in the metric space defined in

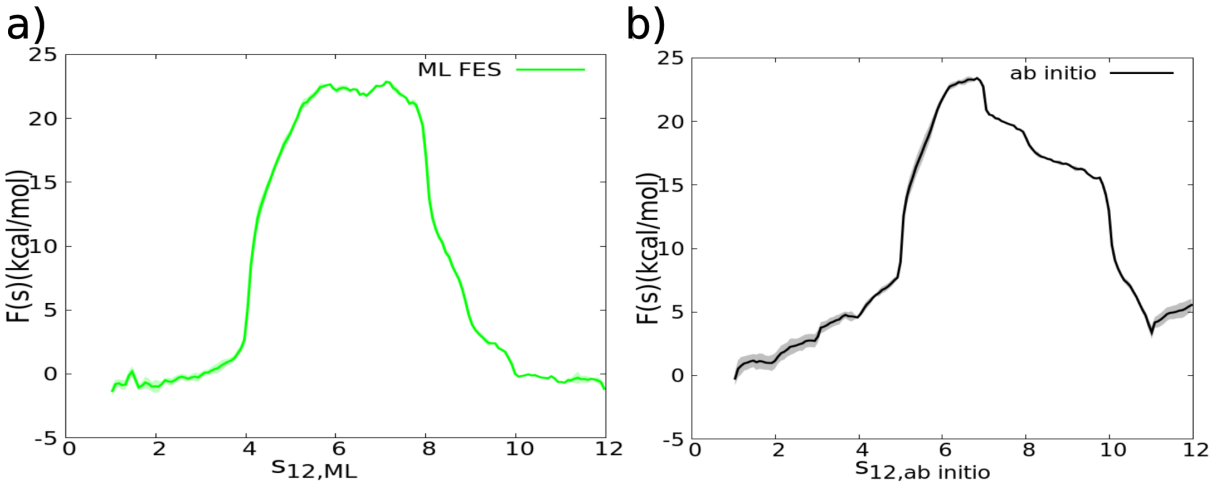

Figure 6.7: Machine learning free energy compared with an *ab intio* free energy along a path collective variable. Panel **a)** free energy from machine learning umbrella sampling simulations along a path collective variable devised with machine learning transition path sampling reference frames (s_{MLP}), panel **b)** the free energy obtained using *ab initio* umbrella sampling simulations along a path collective variable devised with *ab initio* transition path sampling reference frames ($s_{abinitio}$)

equation 1.67, and θ_k is the matrix angle between consecutive path segment. The first term favors equidistant frames, while the second term tends to reduce the length of the chain by favoring a low curvature. After the algorithm is converged, two more references are added at the start and at the end of the path by linearly extrapolating the coordination numbers of the last and first segments to keep metastable states from appearing as spikes in the free-energy landscape: the results are PCV (s, z) based on 12 reference structures.

The results of the machine learning umbrella sampling simulations and the *ab initio* ones are presented in figure 6.7. In the same fashion as for the previous section, a parabolic restraint was set on the chlorine-chlorine coordination number to ensure stability but also on the z collective variable. As the reference frames differ in the $s_{12,ML}$ and $s_{12,abinitio}$, we have decided to display the results in two distinct panels. It is noteworthy that the ML free energy successfully reproduces a precise activation barrier. Furthermore, the profile is more symmetric, and the difference in free energy between the reactants and the products is almost negligible. Initial 60 trajectories were performed however, we notice an insufficient sampling in the vicinity of $s_{12,ML} = 8$, as shown in figure 6.8, this is why we chose to add more sampling windows in that region with a stronger spring constant.

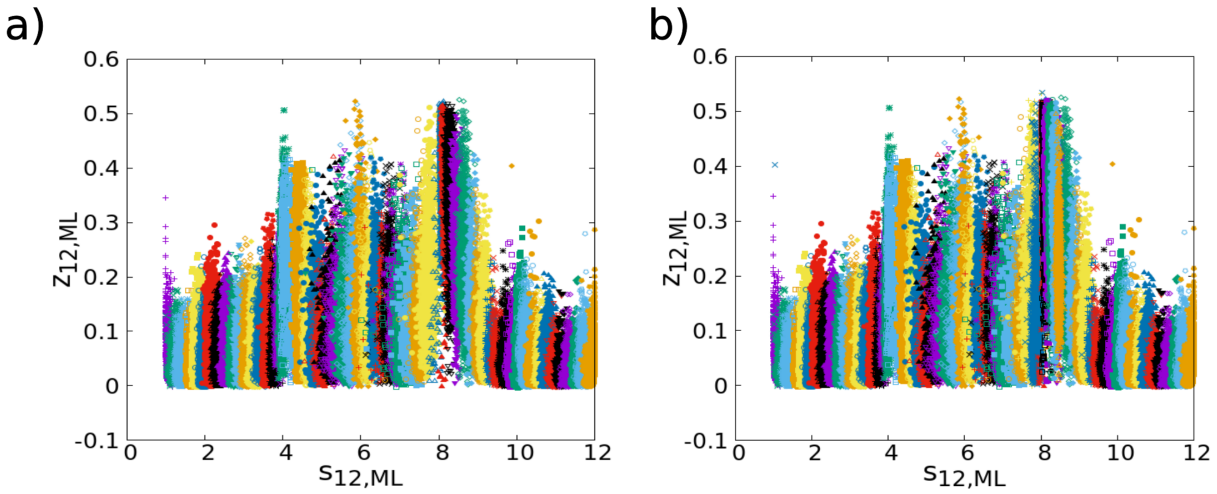

Figure 6.8: Umbrella sampling points from simulations on $s_{12,ML}$ represented on the $(s_{12,ML}, z_{12,ML})$. Panel **a)**, initial sampling on the 60 windows, Panel **b)** sampling with the added windows

6.5 Computation of kinetic rates

We have shown with our MLP that thermodynamic information can be gathered with *ab initio* accuracy for a lesser computational cost. The second big challenge in the study of chemical reactions is the computation of the rate constant. However, it is rarely computed in *ab initio* studies because of the computational burden of the exact transition rate taking into account recrossing events as explained in section 5.2

6.5.1 The transmission coefficient

In section 5.2 we presented the analytical way of obtaining the kinetic rate in equation 5.19, this can be re-written in terms of quantities that are computed using molecular dynamics simulations. Indeed, It can be computationally computed by having N trajectories starting around the top of the barrier and knowing the activation barrier using the following equation:

$$k^{RF}(t) = \frac{\sum_{i=1}^{N} \dot{q}_i(0) h_B(q_i(t))}{N} \frac{e^{-\beta F(q^*)}}{\int_{\Omega_A} e^{-\beta F(q)} dq} \tag{6.2}$$

Where Ω_A is the reactants domain and F is the free energy profile along the chosen CV. The first term is expensive to compute as it needs many trajectories to reach statistical convergence, this is why it is often discarded and transition state theory (TST) is preferred to compute transition rates. The computation of the rate constant includes two terms: the exponential barrier term, and the average over many trajectories, this creates two sources of error for the MLP. Hence, to validate the dynamical behavior of the MLP, we will first compute the transmission coefficient, κ which also characterizes the deviation of

	Machine learning	*ab initio*
κ for $d_1 - d_2$	0.498 ± 0.009	0.54 ± 0.05
κ for $s_{12,ML}$	0.25 ± 0.02	0.25 ± 0.08

Table 6.1: Transmission coefficients (κ) obtained *ab initio* TPS and with machine learning TPS. As the reaction is symmetric, the average on both way was taken and the uncertainty was assessed by computing κ in both halves of the trajectories dataset and taking the deviation as the uncertainty. Around 500 TPS simulations were performed *ab initio* while 5000 were performed with the machine learning potential

the system from transition state theory.

$$\kappa(t) = \frac{k^{RF}(t)}{k^{TST}} = \frac{\sum_{i=1}^{N} \dot{q}_i(0) h_B(q_i(t))}{\sum_{i=1}^{N} \theta(q - q^*)\dot{q}_i(0)} \tag{6.3}$$

κ can also be used to assess the quality of a CV: a very accurate CV will have a TC close to one, while a bad CV will have TC indistinguishable from zero. To compute the rate constant, and the transmission coefficient, we performed around 5000 TPS simulations with our MLP.

The values of κ are reported in table 6.1, the values of the MLP computed κ and the *ab initio* one are within error bars. Which means that the dynamical behavior of our MLP is consistent with the *ab initio* behavior. Moreover, the values obtained here are close to the one given in ref [186] of 0.39 ± 0.07 for $d_1 - d_2$ which confirms the strength of our method. The small discrepancy between our results and the one of reference [186] could be explained by the difference in the electronic structure method. The values of κ also allow us to compare the quality of the CV.

Here, κ is almost twice as high for $d_1 - d_2$ than for $s_{12,ML}$ which means that $d_1 - d_2$ should be a better collective variable than the PCV. This is also the conclusion we reached in reference [24] by using a criterion based on the likelihood of the committor.

We report the most important experimentally measurable quantities in table 6.2. The 1-2kcal/mol (2 k_BT) difference between the two barriers is enough to explain the discrepancy between the rate values, because of its exponential relation with the barrier height. Moreover, the *ab initio* data and the MLP data are in close agreement and the difference between the values reported in this work and the ones from other numerical studies can be explained by the difference in electronic structure calculation methods (The difference in the exchange-correlation functional used or the QM/MM method). The *ab initio* values related to the $s_{12,ML}$ CV were not computed because this would require the free energy profile along this CV which is expensive to compute. This once again illustrates that with this MLP, it is possible to explore free energy profiles along different collective variables, which is impossible with *ab initio* calculations.

Overall, performing a set of US simulations using a MLP has a computational cost of around 50k CPU.h while performing it *ab initio* has a cost of about an order of magnitude

	Machine learning	*ab initio*
K^{eq} for $d_1 - d_2$	1.0 ± 0.8	0.01 ± 0.03
K^{eq} for s_{12}	0.4 ± 0.4	
$\Delta F^{\ddagger}$ for $d_1 - d_2$	20.3 ± 1 kcal/mol	21.9 ± 1 kcal/mol
$\Delta F^{\ddagger}$ for s_{12}	24 kcal/mol	
k for $d_1 - d_2$	$(6 \pm 5) \times 10^{-3} s^{-1}$	$(3 \pm 1) \times 10^{-4} s^{-1}$
k for $s_{12,ML}$	$(1.0 \pm 0.9) \times 10^{-5} s^{-1}$	

Table 6.2: Relevant thermodynamic and kinetic quantities: the equilibrium constant (K^{eq}) related to the relative stability of reactants and products which here is theoretically 1, the barrier height ($\Delta F^{\ddagger}$) corresponding to the stability of the transition state and kinetic rates(k). For the calculation of the barrier height and the kinetic rate, the free energy profile was symmetrized, since the reactants and products are the same. the uncertainties were computed using the propagation of uncertainties formula

more. The computational burden can thus be put in the generation of short *ab initio* unbiased transition pathway to train a MLP and have a first guess of the transition mechanism.

6.6 Committor analysis

The quality of a collective variable can be assessed by committor analysis. As we have harvested a big amount of *ab initio* transition path sampling data for the training of the machine learning potential, we used them as a benchmark and computed the committor probability on the points near the transition of each trajectory. To do so, we shot $N = 200$ trajectories with random initial velocities chosen from the Boltzmann distribution from each point. The committor probability is estimated by doing:

$$\phi_B \approx \frac{n_B}{N} \tag{6.4}$$

The bigger N, the less statistical uncertainty is found in the committor (see ref [193]). We then binned the CV values, of the TPS trajectories with the corresponding computed committor values, we present the results in figure 6.9. The error bars represent the variation of the committor value in one bin, thus high error bars means a big dispersion in the committor prediction.

An ideal collective variable should follow the committor and one value of the collective variable should be mapped to one value of the committor, the plot of the committor vs. the collective variable should therefore be smooth and, in an ideal case the error bars should only be proportional to the variation of the committor function between the two ends of the bin.

This is almost the case for the variable $d_1 - d_2$ in figure 6.9 which confirms our findings with the transmission coefficients and the likelihood approach of ref [24]. On the other hand, it seems that the collective variable taking into account the solvent degrees of

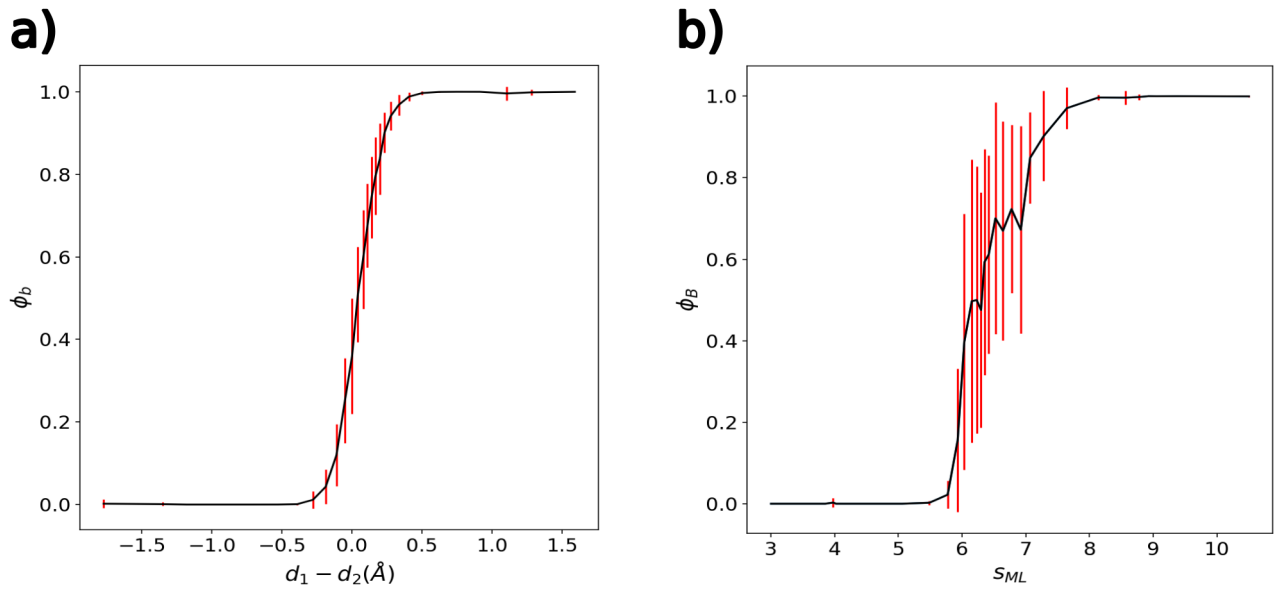

Figure 6.9: Committor function computed along two collective variables. Panel **a)**, $d_1 - d_2$ a heuristic collective variable characterizing the distance of the chlorine ions with respect to the carbon atom. Panel **b)** s_{ML} a path collective variable based on machine learning transition path sampling. The error-bars are computed by binning the data along the collective variable and computing the deviation of the committor value within the bin.

freedom has a worse behavior with respect to the committor. But, this variable has the advantage of being agnostically devised, i.e., this method can be used with any kind of reaction. Furthermore, it was shown that for this particular reaction, the solvent degrees of freedom did not need to be taken into account in the description of the mechanism. They are nonetheless of the utmost importance for the description of the dynamics, indeed, by explicitly describing the solvent, we find a barrier close to the experimental one, while ref [186] report an activation barrier underestimated by 10 kcal/mol by using QM/MM simulations, i.e., only the reactive part and a small shell around are treated using DFT. This encourages the use of a machine learning for future purpose where solvent or environment degrees of freedom are known to take part in the reaction but are too expensive to treat with DFT, an example of these systems is the prebiotic systems.

6.7 Conclusions

In this chapter, we showed that a MLP can be trained using short out of equilibrium trajectories starting from the top of the barrier associated to a rare event. We assess the quality of the training set by performing US simulations with an increasing number of transition path sampling trajectories in the training set. Then, Using US simulations and transition path sampling, we show that this MLP can be used to compute experimentally relevant quantities such as the equilibrium constants and kinetic rates. We applied it to the study of the prototypical S_N2 substitution of methyl-chloride. Finally, we show that short trajectories can be shot from TPS configurations using this MLP to perform

committor analysis to assess the quality of collective variables, which cannot be done using fully *ab initio* calculations due to the computational burden. The final protocol is represented in figure 6.10, it allows more flexibility in the exploration of the mechanism while reducing the computational cost with respect to the one presented in section 1.4.

Moreover, TPS has been used in a variety of applications going from biomolecular systems to nucleation processes, therefore this method paves the way towards the description of complex reactive or rare events phenomena with *ab initio* accuracy with a wide range of applications. For example, the importance of the environment in prebiotic chemistry has already been mentioned in this thesis. The catalytic activity of minerals could be analyzed using this method for bigger simulation boxes.

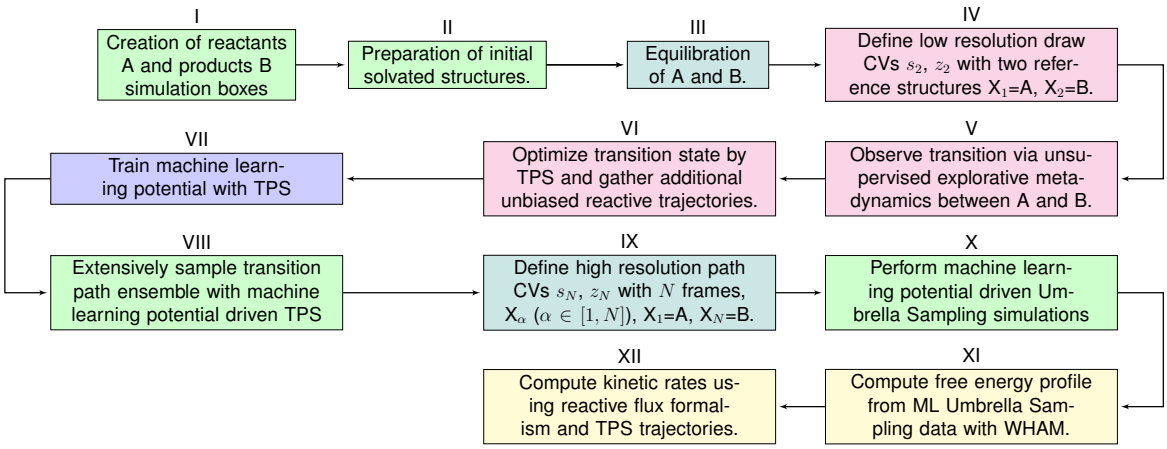

Figure 6.10: Schematic algorithm depicting the simulation protocol. Grey blocks (I, II, IV, IX) indicate pre/post post-proceesing steps where no simulations are needed. Red blocks (III, V, VI) indicate agnostic explorative steps performed using *ab initio* molecular dynamics. Green blocks (VIII, X) indicate quantitative sampling steps performed using machine learning potential. Blue block (VII) indicates the training part of the protocol. The two yellow blocks correspond to the post-processing parts to obtain the relevant thermodynamic and kinetic information from the machine learning simulations

Perspectives and conclusion

During this thesis, we presented two methodologies to study chemical reactions in solution applied to research in origins of life. The first method uses a technique based on a committee to select the parts of the chemical space, where the model does not behave as expected along the reaction coordinate. This allows to target specific parts of the chemical space to put in the training set by performing *ab initio* Umbrella Sampling simulations at the center of these parts. The obtained machine learning potential is used to sample the whole chemical space along the defined reaction coordinate.

This first method comes as an add-on of the already existing protocol of the team, with the addition of the machine learning potential which allows decreasing the computational cost of a study, and therefore to study bigger systems. We applied this new technique to a mechanism of formation of glycine under prebiotic conditions, which contains 8 steps. To the best of our knowledge, this is the first study of a complex reaction network using a machine learning potential. The results obtained are in agreement with the ones obtained in the previous study of the Strecker mechanism. This goes in the favor of our machine learning protocol. Finally, our results are in good agreement with experimental ones. Moreover, they can help interpret the presence of some species in meteorites, such as glycolonitrile or glycolic acid.

The method presented in chapter 3 is however a bit difficult to set into place and requires the prior knowledge of the reaction mechanism and the definition of a collective variable. If the collective variable is bad, and has to be changed, a new machine learning potential will have to be trained.

To overcome this problem, we introduced a method based on transition path sampling to train a machine learning potential. The training process does not depend on the prior knowledge of the transition mechanism, as only a first trajectory bridging the reactants and the products is necessary to generate the training data. This data is generated by shooting trajectories from the top of the barrier and relaxing into the reactants and products basins. With these trajectories, a large amount of configurations between the two wells and the transition state are harvested and are enough to have a machine learning potential accurate on the whole chemical space.

With this trained machine learning potential, it is possible to obtain the free energy profile of the studied mechanism along different collective variables. Furthermore, the kinetic rates can be computed with the reactive flux formalism and compared to experi-

mental results. Finally, the quality of the collective variable can be assessed by performing shooting from transition path sampling configurations via committor analysis. Although this method allows a great speed-up in terms of computational time, there are some limitations to it. Indeed, the training data and thus the TPS simulations still need to be performed which cannot be done if the system is too large.

This approach is novel among the little literature on reactive machine learning force fields in solution. With these two methods, we believe new mechanisms can be studied on top of the one studied in this thesis. For example, in prebiotic chemistry, it is thought that mineral surfaces [194, 195, 196, 197, 198] play a catalytic role in hydrothermal vents and in the interstellar medium, but they are often too expensive to include in simulations with already a solvent and a reactive system. With our methods, we believe it can be studied and this would have a major role in prebiotic chemistry.

On the other hand, in this manuscript, we focused on the case of glycine, which is an amino-acid part of protein that play in two of the three pillars of life defined in the introduction. The replication part is done by storing the information in DNA and RNA molecules, and the question of how they formed on earth is as important as the one of protein formation. Therefore, we also believe that this work might help in studying the mechanisms of formation of building blocks of RNA molecules such as nucleobases and ribose [199, 200].

From a physical point of view, training a potential with transition path sampling data allows harvesting a large amount of different configurations. Another field of dynamics where many configurations appear is the study of nuclear quantum effects. For example, for path integral molecular dynamics, several replicas of the system are created, and a simulation is run for each replica. Since this multiplies the computational time, the electronic degrees of freedom are often treated using less accurate than DFT methods such as density functional based tight binding (DFTB) [201]. Some work has been done to perform path integral molecular dynamics with machine learning potentials [65, 202], but it is rare to see such work for reactive system [203, 204]. We believe this method could help in the study of nuclear quantum effects.

Finally, from a technical point of view, we believe that the equivariant neural network potentials such as nequip and allegro [123, 124] offer a very promising path for the training of even more accurate machine learning potential. They are very new packages and should offer a very attractive alternative to deepmd for reactive force fields.

Appendices

Appendix A

Dissemination of research results and teaching activities

A.1 Publications

A.1.1 Published papers

T. Devergne, T. Magrino, F. Pietrucci, and A. M. Saitta, "Combining Machine Learning Approaches and Accurate Ab Initio Enhanced Sampling Methods for Prebiotic Chemical Reactions in Solution," *J. Chem. Theory Comput.*, vol. 18, pp. 5410–5421, Sept. 2022. Publisher: American Chemical Society.

A.1.2 Papers in preparation

- T. Devergne, L. Huet, F. Pietrucci and A. M. Saitta, "Efficient Machine Learning-based Approach for Accurate Free-Energy Profiles and Kinetic Transition Rates in Chemical Reactions", *In preparation for Phys. Rev. Let.*
- L. Huet, T. Devergne, F. Pietrucci and A. M. Saitta "Machine learning in quantum dynamics for glycine synthesis in origin of life", *In preparation for Proc. Natl. Ac. Soc.*

A.2 Participation to conferences

A.2.1 Organization of conferences/workshop

- Member of the organizing team of "Quantum^2 on machine learning enhanced sampling", 29/10/2023-01/12/2023,CECAM Lausanne, Switzerland,
- Help organization of the "premières journées plénières du GDR IAMAT" (registration), 30/05/2022-01/06/2022, Paris, France

A.2.2 Contributed talk

- Agnostic machine learning description of chemical reactions in solution, *Congrès général des 150 ans de la société française de physique, Mini colloque intelligence artificielle en physique*, 03/07/2023-07/07/2023, Paris, France
- Complete machine learning description of chemical reactions in solution; *American physical society march meeting*, 05/03/2023-10/03/2023, Las Vegas, USA
- Combining machine learning and ab initio enhanced sampling methods for prebiotic chemical reactions, *Conférence exobiologie jeunes chercheurs*, 17/10/2022-19/10/2022, Paris, France
- Combining machine learning and ab initio enhanced sampling methods for prebiotic chemical reactions, *Premières journées plenières GDR IAMAT*, 30/05/2022-01/06/2022, Paris, France

A.2.3 Posters

- Agnostic machine learning description of chemical reactions in solution, *Première école thématique du GDR IAMAT*, 17/04/2023-21/04/2023, Roscoff, France
- Combining machine learning and ab initio enhanced sampling methods for prebiotic chemical reactions, *Machine Learning Meets Statistical Mechanics: Success and Future Challenges in Biosimulations*, 12/10/2022-14/10/2022 CECAM-IT-SIMUL, Grand Hotel Vesuvio, Sorrento, Italy
- Combining machine learning and ab initio enhanced sampling methods for prebiotic chemical reactions, *Chasing CVs using Machine Learning: from methods development to biophysical applications*, 28/06/2022-30/06/2022, CECAM-FR-MOSER Paris, France
- Machine learning and Umbrella sampling to investigate chemical reactions in solution, *Atelier "Méthodes machine-learning pour la modélisation des matériaux", GDR ModMat*,22/09/2021-24/09/2021, Toulouse, France
- Machine learning and Umbrella sampling to investigate chemical reactions in solution, *Paris International school for computational material science*, 30/08/2021-03/09/2021, Paris, France

A.3 Teaching activities

A.3.1 Teaching at the UFR de physique

- Intelligence artificielle pour la physique, Creation, and supervision of practical exercises and supervision on student projects, M1 level (2020 and 2021) (60h)
- Electromagnétisme et ondes, supervision of practicals (optics), L3 level (2021) (10h)

- Structure de la matière, supervision of practicals and problem sets sessions, L3 level (2021) (80h)

A.3.2 Supervision of interns

- Étude du CO_2 géologique à partir des méthodes de Machine Learning, Eliane Farhi, 05/2022-06/2022 (M1 internship)
- Étude des méthodes de machine learning en simulations atomistiques en matière condensée, Mohamed Menshawy, 05/2022-06/2022 (M1 internship)

Bibliography

[1] S. L. Miller, "A Production of Amino Acids Under Possible Primitive Earth Conditions," *Science*, vol. 117, pp. 528–529, May 1953. Publisher: American Association for the Advancement of Science Section: Technical Papers.

[2] A. Laio and M. Parrinello, "Escaping free-energy minima," *PNAS*, vol. 99, pp. 12562–12566, Oct. 2002. Publisher: National Academy of Sciences Section: Physical Sciences.

[3] A. Laio and F. L. Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Rep. Prog. Phys.*, vol. 71, p. 126601, Nov. 2008. Publisher: IOP Publishing.

[4] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, pp. 187–199, Feb. 1977.

[5] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From A to B in free energy space," *J. Chem. Phys.*, vol. 126, p. 054103, Feb. 2007. Publisher: American Institute of Physics.

[6] A. Pérez-Villa, F. Pietrucci, and A. M. Saitta, "Prebiotic chemistry and origins of life research with atomistic computer simulations," *Physics of Life Reviews*, vol. 34-35, pp. 105–135, Dec. 2020.

[7] T. Magrino, F. Pietrucci, and A. M. Saitta, "Step by Step Strecker Amino Acid Synthesis from Ab Initio Prebiotic Chemistry," *J. Phys. Chem. Lett.*, vol. 12, pp. 2630–2637, Mar. 2021. Publisher: American Chemical Society.

[8] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons," *Phys. Rev. Lett.*, vol. 104, p. 136403, Apr. 2010. Publisher: American Physical Society.

[9] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," *Phys. Rev. Lett.*, vol. 98, p. 146401, Apr. 2007. Publisher: American Physical Society.

[10] L. Zhang, J. Han, H. Wang, W. A. Saidi, R. Car, and W. E, "End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems," *arXiv:1805.09003 [cond-mat, physics:physics]*, Dec. 2018. arXiv: 1805.09003.

[11] A. M. Goryaeva, J. Dérès, C. Lapointe, P. Grigorev, T. D. Swinburne, J. R. Kermode, L. Ventelon, J. Baima, and M.-C. Marinica, "Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W," *Phys. Rev. Mater.*, vol. 5, p. 103803, Oct. 2021. Publisher: American Physical Society.

[12] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, "Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics," *J. Phys. Chem. Lett.*, vol. 9, pp. 2879–2885, June 2018. Publisher: American Chemical Society.

[13] T. Morawietz, A. Singraber, C. Dellago, and J. Behler, "How van der Waals interactions determine the unique properties of water," *PNAS*, vol. 113, pp. 8368–8373, July 2016. Publisher: National Academy of Sciences Section: Physical Sciences.

[14] M. Yang, L. Bonati, D. Polino, and M. Parrinello, "Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water," *Catalysis Today*, Mar. 2021.

[15] T. A. Young, T. Johnston-Wood, V. L. Deringer, and F. Duarte, "A transferable active-learning strategy for reactive molecular force fields," *Chem. Sci.*, vol. 12, pp. 10944–10955, Aug. 2021. Publisher: The Royal Society of Chemistry.

[16] T. Devergne, T. Magrino, F. Pietrucci, and A. M. Saitta, "Combining Machine Learning Approaches and Accurate Ab Initio Enhanced Sampling Methods for Prebiotic Chemical Reactions in Solution," *J. Chem. Theory Comput.*, vol. 18, pp. 5410–5421, Sept. 2022. Publisher: American Chemical Society.

[17] G. Kortum, W. Vogel, and K. Andrussow, "Dissociation Constants of Organic Acids in Aqueous Solution."

[18] G. Schlesinger and S. L. Miller, "Equilibrium and kinetics of glyconitrile formation in aqueous solution," *J. Am. Chem. Soc.*, vol. 95, pp. 3729–3735, May 1973. Publisher: American Chemical Society.

[19] J. Jammot, R. Pascal, and A. Commeyras, "Hydration of cyanohydrins in weakly alkaline solutions of boric acid salts," *Tetrahedron Letters*, vol. 30, pp. 563–564, Jan. 1989.

[20] J. Jammot, R. Pascal, and A. Commeyras, "The influence of borate buffers on the hydration rate of cyanohydrins: evidence for an intramolecular mechanism," *J. Chem. Soc., Perkin Trans. 2*, pp. 157–162, Jan. 1990. Publisher: The Royal Society of Chemistry.

[21] P. Haberfield, "What is the energy difference between H2NCH2CO2H and +H3NCH2CO2-?," *J. Chem. Educ.*, vol. 57, p. 346, May 1980. Publisher: American Chemical Society.

[22] V. I. Smirnov and V. G. Badelin, "The enthalpies of solution and solvation of glycine in mixed water-formamide solvents at 298.15 K," *Russ. J. Phys. Chem.*, vol. 80, pp. 357–360, Mar. 2006.

[23] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark," *Annu. Rev. Phys. Chem.*, vol. 53, pp. 291–318, Oct. 2002. Publisher: Annual Reviews.

[24] T. Magrino, L. Huet, A. M. Saitta, and F. Pietrucci, "Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry," *J. Phys. Chem. A*, vol. 126, pp. 8887–8900, Dec. 2022. Publisher: American Chemical Society.

[25] D. Chandler, "Statistical mechanics of isomerization dynamics in liquids and the transition state approximation," *The Journal of Chemical Physics*, vol. 68, no. 6, pp. 2959–2970, 1978.

[26] P. W. Anderson, "More Is Different," *Science*, vol. 177, pp. 393–396, Aug. 1972. Publisher: American Association for the Advancement of Science.

[27] T. Devergne, A. Cattaneo, F. Bournaud, I. Koutsouridou, A. Winter, P. Dimauro, G. A. Mamon, W. Vacher, and M. Varin, "Bulge formation through disc instability - I. Stellar discs," *A&A*, vol. 644, p. A56, Dec. 2020. Publisher: EDP Sciences.

[28] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, "Statistical mechanics for natural flocks of birds," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 4786–4791, Mar. 2012. Publisher: Proceedings of the National Academy of Sciences.

[29] V. J. Kok, M. K. Lim, and C. S. Chan, "Crowd behavior analysis: A review where physics meets biology," *Neurocomputing*, vol. 177, pp. 342–362, Feb. 2016.

[30] L. Caballero, B. Hodge, and S. Hernandez, "Conway's "Game of Life" and the Epigenetic Principle," *Frontiers in Cellular and Infection Microbiology*, vol. 6, 2016.

[31] P. Rendell, "A Universal Turing Machine in Conway's Game of Life," in *2011 International Conference on High Performance Computing & Simulation*, pp. 764–772, July 2011.

[32] S. I. Walker and P. C. W. Davies, "The algorithmic origins of life," *Journal of The Royal Society Interface*, vol. 10, p. 20120869, Feb. 2013. Publisher: Royal Society.

[33] N. Kitadai and S. Maruyama, "Origins of building blocks of life: A review," *Geoscience Frontiers*, vol. 9, Aug. 2017.

[34] S. Bartlett and M. L. Wong, "Defining Lyfe in the Universe: From Three Privileged Functions to Four Pillars," *Life*, vol. 10, p. 42, Apr. 2020. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[35] C. Darwin, "Letter to Joseph Hooker," 1871.

[36] A. Oparin, "The Origin of Life on the Earth - 1st Edition," Aug. 1957.

[37] S. Tirard, "J. B. S. Haldane and the origin of life," *J Genet*, vol. 96, pp. 735–739, Nov. 2017.

[38] M. Theophilo, "Enzymes are Proteins that Act as Biological Catalysts and its Classification and Structure," *Biochemistry & Molecular Biology Journal*, vol. 9, pp. 1–1, Apr. 2023. Publisher: Prime Scholars.

[39] A. Strecker, "Ueber einen neuen aus Aldehyd - Ammoniak und Blausäure entstehenden Körper," *Justus Liebigs Annalen der Chemie*, vol. 91, no. 3, pp. 349–351, 1854. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jlac.18540910309.

[40] J. Oró and S. S. Kamat, "Amino-acid Synthesis from Hydrogen Cyanide under Possible Primitive Earth Conditions," *Nature*, vol. 190, pp. 442–443, Apr. 1961. Number: 4774 Publisher: Nature Publishing Group.

[41] J. Oró and A. P. Kimball, "Synthesis of purines under possible primitive earth conditions. I. Adenine from hydrogen cyanide," *Archives of Biochemistry and Biophysics*, vol. 94, pp. 217–227, Aug. 1961.

[42] E. T. Parker, H. J. Cleaves, J. P. Dworkin, D. P. Glavin, M. Callahan, A. Aubrey, A. Lazcano, and J. L. Bada, "Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, p. 5526, Apr. 2011. Publisher: National Academy of Sciences.

[43] E. T. Parker, J. H. Cleaves, A. S. Burton, D. P. Glavin, J. P. Dworkin, M. Zhou, J. L. Bada, and F. M. Fernández, "Conducting Miller-Urey Experiments," *J Vis Exp*, p. 51039, Jan. 2014.

[44] H. J. Cleaves, "Prebiotic Chemistry: Geochemical Context and Reaction Screening," *Life*, vol. 3, pp. 331–345, June 2013. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[45] W. M. Napier, J. T. Wickramasinghe, and N. C. Wickramasinghe, "The origin of life in comets," *International Journal of Astrobiology*, vol. 6, pp. 321–323, Oct. 2007. Publisher: Cambridge University Press.

[46] A. Bar-Nun, N. Bar-Nun, S. H. Bauer, and C. Sagan, "Shock Synthesis of Amino Acids in Simulated Primitive Environments," *Science*, vol. 168, pp. 470–472, Apr. 1970. Publisher: American Association for the Advancement of Science Section: Reports.

[47] A. S. Burton, J. C. Stern, J. E. Elsila, D. P. Glavin, and J. P. Dworkin, "Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites," *Chem. Soc. Rev.*, vol. 41, pp. 5459–5472, July 2012. Publisher: The Royal Society of Chemistry.

[48] P. d. Marcellus, C. Meinert, I. Myrgorodska, L. Nahon, T. Buhse, L. L. S. d'Hendecourt, and U. J. Meierhenrich, "Aldehydes and sugars from evolved precometary ice analogs: Importance of ices in astrochemical and prebiotic evolution," *PNAS*, vol. 112, pp. 965–970, Jan. 2015. Publisher: National Academy of Sciences Section: Physical Sciences.

[49] W. Martin, J. Baross, D. Kelley, and M. J. Russell, "Hydrothermal vents and the origin of life," *Nat Rev Microbiol*, vol. 6, pp. 805–814, Nov. 2008. Number: 11 Publisher: Nature Publishing Group.

[50] G. Wächtershäuser, "Groundworks for an evolutionary biochemistry: The iron-sulphur world," *Progress in Biophysics and Molecular Biology*, vol. 58, pp. 85–201, Jan. 1992.

[51] A. Brack, "Chapter 10.4 - Clay Minerals and the Origin of Life," in *Developments in Clay Science* (F. Bergaya and G. Lagaly, eds.), vol. 5 of *Handbook of Clay Science*, pp. 507–521, Elsevier, Jan. 2013.

[52] J. P. Ferris, "Mineral Catalysis and Prebiotic Synthesis: Montmorillonite-Catalyzed Formation of RNA," *Elements*, vol. 1, pp. 145–149, June 2005. Publisher: GeoScienceWorld.

[53] P. A. M. Dirac, "Quantum mechanics of many-electron systems," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 123, pp. 714–733, Apr. 1929. Publisher: Royal Society.

[54] A. M. Saitta and F. Saija, "Miller experiments in atomistic computer simulations," *PNAS*, vol. 111, pp. 13768–13773, Sept. 2014. Publisher: National Academy of Sciences Section: Physical Sciences.

[55] R. Spezia, Y. Jeanvoine, W. L. Hase, K. Song, and A. Largo, "Synthesis of formamide and related organic species in the interstellar medium via chemical dynamics simulations," *ApJ*, vol. 826, p. 107, July 2016. Publisher: The American Astronomical Society.

[56] S. Ferrero, S. Pantaleone, C. Ceccarelli, P. Ugliengo, M. Sodupe, and A. Rimola, "Where Does the Energy Go during the Interstellar NH $_3$ Formation on Water Ice? A Computational Study," *ApJ*, vol. 944, p. 142, Feb. 2023.

[57] Y. A. Jeilani and M. T. Nguyen, "Autocatalysis in Formose Reaction and Formation of RNA Nucleosides," *J. Phys. Chem. B*, vol. 124, pp. 11324–11336, Dec. 2020. Publisher: American Chemical Society.

[58] S. Laporte, *The Electric Field at an Oxide Surface - Impact on Reactivity of Prebiotic Molecules*. phdthesis, Université Pierre et Marie Curie - Paris VI, Sept. 2016.

[59] F. Pietrucci, J. C. Aponte, R. Starr, A. Pérez-Villa, J. E. Elsila, J. P. Dworkin, and A. M. Saitta, "Hydrothermal Decomposition of Amino Acids and Origins of Prebiotic Meteoritic Organic Compounds," *ACS Earth Space Chem*, vol. 2, pp. 588–598, Apr. 2018.

[60] A. Pérez-Villa, A. M. Saitta, T. Georgelin, J.-F. Lambert, F. Guyot, M.-C. Maurel, and F. Pietrucci, "Synthesis of RNA Nucleotides in Plausible Prebiotic Conditions from ab Initio Computer Simulations," *J. Phys. Chem. Lett.*, vol. 9, pp. 4981–4987, Sept. 2018. Publisher: American Chemical Society.

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

[62] H. I. Ingólfsson, C. Arnarez, X. Periole, and S. J. Marrink, "Computational 'microscopy' of cellular membranes," *Journal of Cell Science*, vol. 129, pp. 257–268, Jan. 2016.

[63] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, Aug. 2021. Number: 7873 Publisher: Nature Publishing Group.

[64] A. V. Shapeev, "Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials," *Multiscale Model. Simul.*, vol. 14, pp. 1153–1173, Jan. 2016. Publisher: Society for Industrial and Applied Mathematics.

[65] J. Daru, H. Forbert, J. Behler, and D. Marx, "Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark," *Phys. Rev. Lett.*, vol. 129, p. 226001, Nov. 2022. Publisher: American Physical Society.

[66] J. Chandrasekhar, S. F. Smith, and W. L. Jorgensen, "Theoretical examination of the SN2 reaction involving chloride ion and methyl chloride in the gas phase and aqueous solution," *J. Am. Chem. Soc.*, vol. 107, pp. 154–163, Jan. 1985. Publisher: American Chemical Society.

[67] D. R. Hartree, "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 89–110, Jan. 1928. Publisher: Cambridge University Press.

[68] V. Fock, "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems," *Z. Physik*, vol. 61, pp. 126–148, Jan. 1930.

[69] M. Born and R. Oppenheimer, "Zur Quantentheorie der Molekeln," *Annalen der Physik*, vol. 389, no. 20, pp. 457–484, 1927. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19273892002.

[70] J. C. Slater, "The Theory of Complex Spectra," *Phys. Rev.*, vol. 34, pp. 1293–1322, Nov. 1929. Publisher: American Physical Society.

[71] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas," *Phys. Rev.*, vol. 136, pp. B864–B871, Nov. 1964. Publisher: American Physical Society.

[72] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," *Phys. Rev.*, vol. 140, pp. A1133–A1138, Nov. 1965. Publisher: American Physical Society.

[73] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*. Cambridge: Cambridge University Press, 2004.

[74] D. M. Ceperley and B. J. Alder, "Ground State of the Electron Gas by a Stochastic Method," *Phys. Rev. Lett.*, vol. 45, pp. 566–569, Aug. 1980. Publisher: American Physical Society.

[75] D. Bagayoko, "Understanding density functional theory (DFT) and completing it in practice," *AIP Advances*, vol. 4, p. 127104, Dec. 2014.

[76] G. L. Zhao, D. Bagayoko, and T. D. Williams, "Local-density-approximation prediction of electronic properties of GaN, Si, C, and ${\mathrm{RuO}}_{2}$," *Phys. Rev. B*, vol. 60, pp. 1563–1572, July 1999. Publisher: American Physical Society.

[77] C. Lee, D. Vanderbilt, K. Laasonen, R. Car, and M. Parrinello, "Ab initio studies on the structural and dynamical properties of ice," *Phys. Rev. B*, vol. 47, pp. 4863–4872, Mar. 1993. Publisher: American Physical Society.

[78] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Phys. Rev. Lett.*, vol. 77, pp. 3865–3868, Oct. 1996. Publisher: American Physical Society.

[79] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *The Journal of Chemical Physics*, vol. 98, pp. 5648–5652, Apr. 1993.

[80] J. Moellmann and S. Grimme, "DFT-D3 Study of Some Molecular Crystals," *J. Phys. Chem. C*, vol. 118, pp. 7615–7621, Apr. 2014. Publisher: American Chemical Society.

[81] P. Hai and C. Wu, "A comparative DFT study of the oxidation of Al crystals and nanoparticles," *Phys. Chem. Chem. Phys.*, vol. 23, pp. 24004–24015, Oct. 2021. Publisher: The Royal Society of Chemistry.

[82] J. Wu, "Density Functional Theory for Liquid Structure and Thermodynamics," in *Molecular Thermodynamics of Complex Systems* (X. Lu and Y. Hu, eds.), Structure and Bonding, pp. 1–73, Berlin, Heidelberg: Springer, 2009.

[83] N. Troullier and J. L. Martins, "Efficient pseudopotentials for plane-wave calculations," *Phys. Rev. B*, vol. 43, pp. 1993–2006, Jan. 1991. Publisher: American Physical Society.

[84] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," *AIP Conference Proceedings*, vol. 577, pp. 1–20, July 2001.

[85] L. Verlet, "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules," *Phys. Rev.*, vol. 159, pp. 98–103, July 1967. Publisher: American Physical Society.

[86] S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," *The Journal of Chemical Physics*, vol. 81, pp. 511–519, July 1984.

[87] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Phys. Rev. A*, vol. 31, pp. 1695–1697, Mar. 1985. Publisher: American Physical Society.

[88] J. Zielkewicz, "Structural properties of water: Comparison of the SPC, SPCE, TIP4P, and TIP5P models of water | The Journal of Chemical Physics | AIP Publishing," 2005.

[89] P. Mark and L. Nilsson, "Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K," *J. Phys. Chem. A*, vol. 105, pp. 9954–9960, Nov. 2001. Publisher: American Chemical Society.

[90] H. Eyring, "The Activated Complex in Chemical Reactions," *The Journal of Chemical Physics*, vol. 3, pp. 107–115, Feb. 1935.

[91] P. Tiwary and M. Parrinello, "From Metadynamics to Dynamics," *Phys. Rev. Lett.*, vol. 111, p. 230602, Dec. 2013. Publisher: American Physical Society.

[92] B. Roux, "The calculation of the potential of mean force using computer simulations," *Computer Physics Communications*, vol. 91, pp. 275–282, Sept. 1995.

[93] A. L. Ferguson, "BayesWHAM: A Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method," *Journal of Computational Chemistry*, vol. 38, no. 18, pp. 1583–1605, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.24800.

[94] F. Zhu and G. Hummer, "Convergence and error estimation in free energy calculations using the weighted histogram analysis method," *Journal of Computational Chemistry*, vol. 33, no. 4, pp. 453–465, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21989.

[95] E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy, "Temperature Weighted Histogram Analysis Method, Replica Exchange, and Transition Paths," *J. Phys. Chem. B*, vol. 109, pp. 6722–6731, Apr. 2005. Publisher: American Chemical Society.

[96] A. Grossfield, "WHAM: the weighted histogram analysis method version 2.0.10.2."

[97] F. Pietrucci and A. M. Saitta, "Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios," *PNAS*, vol. 112, pp. 15030–15035, Dec. 2015. Publisher: National Academy of Sciences Section: Physical Sciences.

[98] L. Bottou, "Stochastic Gradient Descent Tricks," in *Neural Networks: Tricks of the Trade: Second Edition* (G. Montavon, G. B. Orr, and K.-R. Müller, eds.), Lecture Notes in Computer Science, pp. 421–436, Berlin, Heidelberg: Springer, 2012.

[99] C. M. Bishop, "Training with Noise is Equivalent to Tikhonov Regularization," *Neural Computation*, vol. 7, pp. 108–116, Jan. 1995.

[100] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima," Tech. Rep. arXiv:1609.04836, arXiv, Feb. 2017. arXiv:1609.04836 [cs, math] type: article.

[101] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017. arXiv: 1412.6980.

[102] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, Dec. 1943.

[103] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, pp. 386–408, 1958. Place: US Publisher: American Psychological Association.

[104] M. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, Sept. 2017.

[105] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986. Number: 6088 Publisher: Nature Publishing Group.

[106] CNRS - Formation FIDLE, "Fidle - Séquence 01 (Live)," Nov. 2021.

[107] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, pp. 551–560, Jan. 1990.

[108] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signal Systems*, vol. 2, pp. 303–314, Dec. 1989.

[109] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, "Neural network models of potential energy surfaces," *The Journal of Chemical Physics*, vol. 103, pp. 4129–4137, Sept. 1995.

[110] S. Manzhos and T. Carrington, Jr., "Using redundant coordinates to represent potential energy surfaces with lower-dimensional functions," *The Journal of Chemical Physics*, vol. 127, p. 014103, July 2007.

[111] M. Malshe, R. Narulkar, L. M. Raff, M. Hagan, S. Bukkapatnam, P. M. Agrawal, and R. Komanduri, "Development of generalized potential-energy surfaces using many-body expansions, neural networks, and moiety energy approximations," *The Journal of Chemical Physics*, vol. 130, p. 184102, May 2009.

[112] S. Hobday, R. Smith, and J. Belbruno, "Applications of neural networks to fitting interatomic potential functions," *Modelling Simul. Mater. Sci. Eng.*, vol. 7, p. 397, May 1999.

[113] J. Behler, "Constructing high-dimensional neural network potentials: A tutorial review," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1032–1050, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24890.

[114] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 13754–13769, May 2016. Publisher: The Royal Society of Chemistry.

[115] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1094–1101, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24917.

[116] H. Huo and M. Rupp, "Unified Representation of Molecules and Crystals for Machine Learning," *Mach. Learn.: Sci. Technol.*, vol. 3, p. 045017, Dec. 2022. arXiv:1704.06439 [cond-mat, physics:physics].

[117] S. Bhakat, "Collective variable discovery in the age of machine learning: reality, hype and everything in between," *RSC Adv.*, vol. 12, pp. 25010–25024, Aug. 2022. Publisher: The Royal Society of Chemistry.

[118] M. Moog, *Carbon dioxide at extreme conditions : liquid(s), crystals, glasses and their transformation from ab initio topological methods.* These de doctorat, Sorbonne université, Sept. 2019.

[119] M. Moog, F. Pietrucci, and A. M. Saitta, "Carbon Dioxide under Earth Mantle Conditions: From a Molecular Liquid through a Reactive Fluid to Polymeric Regimes," *J. Phys. Chem. A*, vol. 125, pp. 5863–5869, July 2021. Publisher: American Chemical Society.

[120] H. Wang, L. Zhang, J. Han, and W. E, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics," *Computer Physics Communications*, vol. 228, pp. 178–184, July 2018.

[121] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, and W. E, "DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models," *Computer Physics Communications*, vol. 253, p. 107206, Aug. 2020.

[122] D. Lu, H. Wang, M. Chen, L. Lin, R. Car, W. E, W. Jia, and L. Zhang, "86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy," *Computer Physics Communications*, vol. 259, p. 107624, Feb. 2021.

[123] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," *Nat Commun*, vol. 13, p. 2453, May 2022. Number: 1 Publisher: Nature Publishing Group.

[124] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky, "Learning local equivariant representations for large-scale atomistic dynamics," *Nat Commun*, vol. 14, p. 579, Feb. 2023. Number: 1 Publisher: Nature Publishing Group.

[125] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B*, vol. 87, p. 184115, May 2013. Publisher: American Physical Society.

[126] R. K. Lindsey, L. E. Fried, and N. Goldman, "ChIMES: A Force Matched Potential with Explicit Three-Body Interactions for Molten Carbon," *J. Chem. Theory Comput.*, vol. 13, pp. 6222–6229, Dec. 2017. Publisher: American Chemical Society.

[127] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, and K. Burke, "Quantum chemical accuracy from density functional approximations via machine learning," *Nat Commun*, vol. 11, p. 5223, Oct. 2020. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational chemistry;Computational science Subject_term_id: computational-chemistry;computational-science.

[128] N. Gerrits, K. Shakouri, J. Behler, and G.-J. Kroes, "Accurate Probabilities for Highly Activated Reaction of Polyatomic Molecules on Surfaces Using a High-Dimensional Neural Network Potential: CHD3 + Cu(111)," *J Phys Chem Lett*, vol. 10, pp. 1763–1768, Apr. 2019.

[129] C. Schran, K. Brezina, and O. Marsalek, "Committee neural network potentials control generalization errors and enable active learning," *J. Chem. Phys.*, vol. 153, p. 104105, Sept. 2020. Publisher: American Institute of Physics.

[130] N. Xu, Y. Shi, Y. He, and Q. Shao, "A Deep-Learning Potential for Crystalline and Amorphous Li–Si Alloys," *J. Phys. Chem. C*, p. 11, 2020.

[131] L. Zhang, H. Wang, R. Car, and W. E, "Phase Diagram of a Deep Potential Water Model," *Phys. Rev. Lett.*, vol. 126, p. 236001, June 2021. Publisher: American Physical Society.

[132] G. Imbalzano, Y. Zhuang, V. Kapil, K. Rossi, E. A. Engel, F. Grasselli, and M. Ceriotti, "Uncertainty estimation for molecular dynamics and sampling," *J. Chem. Phys.*, vol. 154, p. 074102, Feb. 2021. Publisher: American Institute of Physics.

[133] C. Schran, F. L. Thiemann, P. Rowe, E. A. Muller, O. Marsalek, and A. Michaelides, "Machine learning potentials for complex aqueous systems made simple," p. 20.

[134] L. Kahle and F. Zipoli, "On the Quality of Uncertainty Estimates from Neural Network Potential Ensembles," *Phys. Rev. E*, vol. 105, p. 015311, Jan. 2022. arXiv: 2108.05748.

[135] J. Xu, X.-M. Cao, and P. Hu, "Accelerating Metadynamics-Based Free-Energy Calculations with Adaptive Machine Learning Potentials," *J. Chem. Theory Comput.*, June 2021. Publisher: American Chemical Society.

[136] X. Pan, J. Yang, R. Van, E. Epifanovsky, J. Ho, J. Huang, J. Pu, Y. Mei, K. Nam, and Y. Shao, "Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions," *J. Chem. Theory Comput.*, vol. 17, pp. 5745–5758, Sept. 2021. Publisher: American Chemical Society.

[137] S. Plimpton, "Fast Parallel Algorithms for Short-Range Molecular Dynamics," *Journal of Computational Physics*, vol. 117, pp. 1–19, Mar. 1995.

[138] M. Bonomi, G. Bussi, C. Camilloni, G. A. Tribello, P. Banáš, A. Barducci, M. Bernetti, P. G. Bolhuis, S. Bottaro, D. Branduardi, R. Capelli, P. Carloni, M. Ceriotti, A. Cesari,

H. Chen, W. Chen, F. Colizzi, S. De, M. De La Pierre, D. Donadio, V. Drobot, B. Ensing, A. L. Ferguson, M. Filizola, J. S. Fraser, H. Fu, P. Gasparotto, F. L. Gervasio, F. Giberti, A. Gil-Ley, T. Giorgino, G. T. Heller, G. M. Hocky, M. Iannuzzi, M. Invernizzi, K. E. Jelfs, A. Jussupow, E. Kirilin, A. Laio, V. Limongelli, K. Lindorff-Larsen, T. Löhr, F. Marinelli, L. Martin-Samos, M. Masetti, R. Meyer, A. Michaelides, C. Molteni, T. Morishita, M. Nava, C. Paissoni, E. Papaleo, M. Parrinello, J. Pfaendtner, P. Piaggi, G. Piccini, A. Pietropaolo, F. Pietrucci, S. Pipolo, D. Provasi, D. Quigley, P. Raiteri, S. Raniolo, J. Rydzewski, M. Salvalaglio, G. C. Sosso, V. Spiwok, J. Šponer, D. W. H. Swenson, P. Tiwary, O. Valsson, M. Vendruscolo, G. A. Voth, A. White, and The PLUMED consortium, "Promoting transparency and reproducibility in enhanced molecular simulations," *Nat Methods*, vol. 16, pp. 670–673, Aug. 2019. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Culture;Software Subject_term_id: culture;software.

[139] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "PLUMED 2: New feathers for an old bird," *Computer Physics Communications*, vol. 185, pp. 604–613, Feb. 2014.

[140] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "Multidimensional free-energy calculations using the weighted histogram analysis method," *Journal of Computational Chemistry*, vol. 16, no. 11, pp. 1339–1350, 1995. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540161104.

[141] T. Bučko, S. Chibani, J.-F. Paul, L. Cantrel, and M. Badawi, "Dissociative iodomethane adsorption on Ag-MOR and the formation of AgI clusters: an ab initio molecular dynamics study," *Phys. Chem. Chem. Phys.*, vol. 19, pp. 27530–27543, Oct. 2017. Publisher: The Royal Society of Chemistry.

[142] S. Pizzarello, "The Chemistry of Life's Origin: A Carbonaceous Meteorite Perspective," *Acc. Chem. Res.*, vol. 39, pp. 231–237, Apr. 2006. Publisher: American Chemical Society.

[143] S. Pizzarello, Y. Huang, and M. R. Alexandre, "Molecular asymmetry in extraterrestrial chemistry: Insights from a pristine meteorite," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 3700–3704, Mar. 2008. Publisher: Proceedings of the National Academy of Sciences.

[144] J. E. Elsila, D. P. Glavin, and J. P. Dworkin, "Cometary glycine detected in samples returned by Stardust," *Meteoritics & Planetary Science*, vol. 44, no. 9, pp. 1323–1330, 2009. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1945-5100.2009.tb01224.x.

[145] S. A. Sandford, J. Aléon, C. M. O. Alexander, T. Araki, S. Bajt, G. A. Baratta, J. Borg, J. P. Bradley, D. E. Brownlee, J. R. Brucato, M. J. Burchell, H. Busemann, A. Butterworth, S. J. Clemett, G. Cody, L. Colangeli, G. Cooper, L. D'Hendecourt, Z. Djouadi, J. P. Dworkin, G. Ferrini, H. Fleckenstein, G. J. Flynn, I. A. Franchi, M. Fries, M. K. Gilles, D. P. Glavin, M. Gounelle, F. Grossemy, C. Jacobsen, L. P. Keller, A. L. D. Kilcoyne, J. Leitner, G. Matrajt, A. Meibom, V. Mennella, S. Mostefaoui, L. R. Nittler, M. E. Palumbo, D. A. Papanastassiou, F. Robert, A. Rotundi, C. J. Snead, M. K. Spencer, F. J. Stadermann, A. Steele, T. Stephan, P. Tsou, T. Tyliszczak, A. J. Westphal, S. Wirick, B. Wopenka, H. Yabuta, R. N. Zare, and M. E. Zolensky, "Organics Captured from Comet 81P/Wild 2 by the Stardust Spacecraft," *Science*, vol. 314, pp. 1720–1724, Dec. 2006. Publisher: American Association for the Advancement of Science.

[146] Z. Martins, "Organic Chemistry of Carbonaceous Meteorites," *Elements*, vol. 7, pp. 35–40, Feb. 2011. Publisher: GeoScienceWorld.

[147] C. Ceccarelli, L. Loinard, A. Castets, A. Faure, and B. Lefloch, "Search for glycine in the solar type protostar IRAS 16293-2422," *Astronomy and Astrophysics*, vol. 362, pp. 1122–1126, Sept. 2000.

[148] L. E. Snyder, F. J. Lovas, J. M. Hollis, D. N. Friedel, P. R. Jewell, A. Remijan, V. V. Ilyushin, E. A. Alekseev, and S. F. Dyubko, "A Rigorous Attempt to Verify Interstellar Glycine," *ApJ*, vol. 619, p. 914, Feb. 2005. Publisher: IOP Publishing.

[149] L. E. Snyder, "THE SEARCH FOR INTERSTELLAR GLYCINE," *Orig Life Evol Biosph*, vol. 27, pp. 115–133, June 1997.

[150] Y.-J. Kuan, S. B. Charnley, H.-C. Huang, W.-L. Tseng, and Z. Kisiel, "Interstellar Glycine," *ApJ*, vol. 593, p. 848, Aug. 2003. Publisher: IOP Publishing.

[151] G. Moutou, J. Taillades, S. Bénefice-Malouet, A. Commeyras, G. Messina, and R. Mansani, "Equilibrium of alpha-aminoacetonitrile formation from formaldehyde, hydrogen cyanide and ammonia in aqueous solution: Industrial and prebiotic significance," *Journal of Physical Organic Chemistry*, vol. 8, no. 11, pp. 721–730, 1995. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/poc.610081105.

[152] G. Danger, F. Duvernay, P. Theulé, F. Borget, and T. Chiavassa, "HYDROXYACETONITRILE (HOCH $_2$ CN) FORMATION IN ASTROPHYSICAL CONDITIONS. COMPETITION WITH THE AMINOMETHANOL, A GLYCINE PRECURSOR," *ApJ*, vol. 756, p. 11, Sept. 2012.

[153] K. L. Thrush and J. Kua, "Reactions of Glycolonitrile with Ammonia and Water: A Free Energy Map," *J. Phys. Chem. A*, vol. 122, pp. 6769–6779, Aug. 2018. Publisher: American Chemical Society.

[154] R. Arnaud, C. Adamo, M. Cossi, A. Milet, Y. Vallée, and V. Barone, "Theoretical Study of the Addition of Hydrogen Cyanide to Methanimine in the Gas Phase and in Aqueous Solution," *J. Am. Chem. Soc.*, vol. 122, pp. 324–330, Jan. 2000. Publisher: American Chemical Society.

[155] D. E. Woon, "Ab Initio Quantum Chemical Studies of Reactions in Astrophysical Ices: 2. Reactions in H2CO/HCN/HNC/H2O Ices," *Icarus*, vol. 149, pp. 277–284, Jan. 2001.

[156] M. Ferus, F. Pietrucci, A. M. Saitta, O. Ivanek, A. Knizek, P. Kubelík, M. Krus, L. Juha, R. Dudzak, J. Dostál, A. Pastorek, L. Petera, J. Hrncirova, H. Saeidfirozeh, V. Shestivská, J. Sponer, J. E. Sponer, P. Rimmer, S. Civiš, and G. Cassone, "Prebiotic synthesis initiated in formaldehyde by laser plasma simulating high-velocity impacts," *A&A*, vol. 626, p. A52, June 2019. Publisher: EDP Sciences.

[157] L. Margulès, B. A. McGuire, M. L. Senent, R. A. Motiyenko, A. Remijan, and J. C. Guillemin, "Submillimeter spectra of 2-hydroxyacetonitrile (glycolonitrile; HOCH $_2$ CN) and its searches in GBT PRIMOS observations of Sgr B2(N)," *A&A*, vol. 601, p. A50, May 2017.

[158] G. Danger, F. Duvernay, P. Theulé, F. Borget, J.-C. Guillemin, and T. Chiavassa, "Hydroxyacetonitrile (HOCH $_2$ CN) as a precursor for formylcyanide (CHOCN), ketenimine

(CH $_2$ CNH), and cyanogen (NCCN) in astrophysical conditions," *A&A*, vol. 549, p. A93, Jan. 2013.

[159] L. Piani, H. Yurimoto, and L. Remusat, "A dual origin for water in carbonaceous asteroids revealed by CM chondrites," *Nat Astron*, vol. 2, pp. 317–323, Apr. 2018. Number: 4 Publisher: Nature Publishing Group.

[160] M. D. Suttle, A. J. King, P. F. Schofield, H. Bates, and S. S. Russell, "The aqueous alteration of CM chondrites, a review," *Geochimica et Cosmochimica Acta*, vol. 299, pp. 219–256, Apr. 2021.

[161] L. Rotelli, J. M. Trigo-Rodríguez, C. E. Moyano-Cambero, E. Carota, L. Botta, E. Di Mauro, and R. Saladino, "The key role of meteorites in the formation of relevant prebiotic molecules in a formamide/water environment," *Sci Rep*, vol. 6, p. 38888, Dec. 2016.

[162] V. Vinogradoff, S. Bernard, C. Le Guillou, and L. Remusat, "Evolution of interstellar organic compounds under asteroidal hydrothermal conditions," *Icarus*, vol. 305, pp. 358–370, May 2018.

[163] V. Vinogradoff, C. Le Guillou, S. Bernard, L. Binet, P. Cartigny, A. J. Brearley, and L. Remusat, "Paris vs. Murchison: Impact of hydrothermal alteration on organic matter in CM chondrites," *Geochimica et Cosmochimica Acta*, vol. 212, pp. 234–252, Sept. 2017.

[164] C. Le Guillou, S. Bernard, A. J. Brearley, and L. Remusat, "Evolution of organic matter in Orgueil, Murchison and Renazzo during parent body aqueous alteration: In situ investigations," *Geochimica et Cosmochimica Acta*, vol. 131, pp. 368–392, Apr. 2014.

[165] G. W. Cooper and J. R. Cronin, "Linear and cyclic aliphatic carboxamides of the Murchison meteorite: Hydrolyzable derivatives of amino acids and other carboxylic acids," *Geochimica et Cosmochimica Acta*, vol. 59, pp. 1003–1015, Mar. 1995.

[166] G. Cooper, N. Kimmich, W. Belisle, J. Sarinana, K. Brabham, and L. Garrel, "Carbonaceous meteorites as a source of sugar-related organic compounds for the early Earth," *Nature*, vol. 414, pp. 879–883, Dec. 2001. Number: 6866 Publisher: Nature Publishing Group.

[167] J. R. Cronin and S. Chang, "Organic Matter in Meteorites: Molecular and Isotopic Analyses of the Murchison Meteorite," in *The Chemistry of Life's Origins* (J. M. Greenberg, C. X. Mendoza-Gómez, and V. Pirronello, eds.), NATO ASI Series, pp. 209–258, Dordrecht: Springer Netherlands, 1993.

[168] L. Mouaffac, K. Palacio-Rodriguez, and F. Pietrucci, "Optimal reaction coordinates and kinetic rates from the projected dynamics of transition paths," Tech. Rep. arXiv:2302.12497, arXiv, Feb. 2023. arXiv:2302.12497 [cond-mat, physics:physics] type: article.

[169] L. Bonati, G. Piccini, and M. Parrinello, "Deep learning the slow modes for rare events sampling," *Proceedings of the National Academy of Sciences*, vol. 118, p. e2113533118, Nov. 2021. Publisher: Proceedings of the National Academy of Sciences.

[170] R. G. Mullen, J.-E. Shea, and B. Peters, "Easy Transition Path Sampling Methods: Flexible-Length Aimless Shooting and Permutation Shooting," *J. Chem. Theory Comput.*, vol. 11, pp. 2421–2428, June 2015. Publisher: American Chemical Society.

[171] Z. Belkacemi, P. Gkeka, T. Lelièvre, and G. Stoltz, "Chasing Collective Variables Using Autoencoders and Biased Trajectories," *J. Chem. Theory Comput.*, vol. 18, pp. 59–78, Jan. 2022. Publisher: American Chemical Society.

[172] L. Onsager, "Initial Recombination of Ions," *Phys. Rev.*, vol. 54, pp. 554–557, Oct. 1938. Publisher: American Physical Society.

[173] D. Ryter, "On the eigenfunctions of the Fokker-Planck operator and of its adjoint," *Physica A: Statistical Mechanics and its Applications*, vol. 142, pp. 103–121, Apr. 1987.

[174] A. Berezhkovskii and A. Szabo, "Perturbation theory of phi-value analysis of two-state protein folding: Relation between pfold and phi values," *The Journal of Chemical Physics*, vol. 125, p. 104902, Sept. 2006.

[175] S. Jungblut and C. Dellago, "Pathways to self-organization: Crystallization via nucleation and growth," *Eur. Phys. J. E*, vol. 39, p. 77, Aug. 2016.

[176] W. E, W. Ren, and E. Vanden-Eijnden, "Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes," *Chemical Physics Letters*, vol. 413, pp. 242–247, Sept. 2005.

[177] T. S. van Erp, D. Moroni, and P. G. Bolhuis, "A novel path sampling method for the calculation of rate constants," *The Journal of Chemical Physics*, vol. 118, pp. 7762–7774, Apr. 2003.

[178] R. J. Allen, C. Valeriani, and P. R. t. Wolde, "Forward Flux Sampling for rare event simulations," *J. Phys.: Condens. Matter*, vol. 21, p. 463102, Nov. 2009. arXiv:0906.4758 [cond-mat].

[179] 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, p. 152716, Aug. 2017.

[180] 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, p. 034109, July 2007.

[181] A. Jedrecy, "Study of phase transformation of matter through topological coordinates - TEL - Thèses en ligne."

[182] H. Eyring, "The theory of absolute reaction rates," *Trans. Faraday Soc.*, vol. 34, pp. 41–48, Jan. 1938. Publisher: The Royal Society of Chemistry.

[183] E. Wigner, "The transition state method," *Trans. Faraday Soc.*, vol. 34, pp. 29–41, Jan. 1938. Publisher: The Royal Society of Chemistry.

[184] D. Frenkel and B. Smit, "Chapter 16 - Rare Events," in *Understanding Molecular Simulation (Second Edition)* (D. Frenkel and B. Smit, eds.), pp. 431–464, San Diego: Academic Press, Jan. 2002.

[185] B. Ensing, E. J. Meijer, P. E. Blöchl, and E. J. Baerends, "Solvation Effects on the SN2 Reaction between CH3Cl and Cl- in Water," *J. Phys. Chem. A*, vol. 105, pp. 3300–3310, Apr. 2001. Publisher: American Chemical Society.

[186] C. Leitold, C. J. Mundy, M. D. Baer, G. K. Schenter, and B. Peters, "Solvent reaction coordinate for an SN2 reaction," *The Journal of Chemical Physics*, vol. 153, p. 024103, July 2020.

[187] A. J. Parker, "Protic-dipolar aprotic solvent effects on rates of bimolecular reactions," *Chem. Rev.*, vol. 69, pp. 1–32, Feb. 1969. Publisher: American Chemical Society.

[188] N. M. M. Nibbering, "Mechanistic Aspects of Ion-Molecule Reactions," in *Kinetics of Ion-Molecule Reactions* (P. Ausloos, ed.), NATO Advanced Study Institutes Series, pp. 165–197, Boston, MA: Springer US, 1979.

[189] W. J. Albery and M. M. Kreevoy, "Methyl Transfer Reactions," in *Advances in Physical Organic Chemistry* (V. Gold and D. Bethell, eds.), vol. 16, pp. 87–157, Academic Press, Jan. 1978.

[190] R. L. Heppolette, R. E. Robertson, and E. W. R. Steacie, "The neutral hydrolysis of the methyl halides," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 252, pp. 273–285, Jan. 1997. Publisher: Royal Society.

[191] W. N. Olmstead and J. I. Brauman, "Gas-phase nucleophilic displacement reactions," *J. Am. Chem. Soc.*, vol. 99, pp. 4219–4228, June 1977. Publisher: American Chemical Society.

[192] G. Winter and R. Gómez-Bombarelli, "Simulations with machine learning potentials identify the ion conduction mechanism mediating non-Arrhenius behavior in LGPS," *J. Phys. Energy*, vol. 5, p. 024004, Apr. 2023.

[193] C. Dellago, P. G. Bolhuis, and P. L. Geissler, "Transition Path Sampling," in *Advances in Chemical Physics*, pp. 1–78, John Wiley & Sons, Ltd, 2002. Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471231509.ch1.

[194] A. Aithal, S. Dagar, and S. Rajamani, "Metals in Prebiotic Catalysis: A Possible Evolutionary Pathway for the Emergence of Metalloproteins," *ACS Omega*, vol. 8, pp. 5197–5208, Feb. 2023. Publisher: American Chemical Society.

[195] V. Erastova, M. T. Degiacomi, D. G. Fraser, and H. C. Greenwell, "Mineral surface chemistry control for origin of prebiotic peptides," *Nat Commun*, vol. 8, p. 2033, Dec. 2017. Number: 1 Publisher: Nature Publishing Group.

[196] Y. Li, "Minerals as Prebiotic Catalysts for Chemical Evolution towards the Origin of Life," in *Mineralogy*, IntechOpen, Feb. 2022.

[197] A. Rimola, M. Sodupe, and P. Ugliengo, "Role of Mineral Surfaces in Prebiotic Chemical Evolution. In Silico Quantum Mechanical Studies," *Life*, vol. 9, Mar. 2019. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[198] E. Mateo-Marti, S. Galvez-Martinez, E. Cueto-Diaz, and M. P. Zorzano, "The role of minerals surfaces in prebiotic chemistry and planetary exploration," Tech. Rep. EPSC2022-100, Copernicus Meetings, July 2022. Conference Name: EPSC2022.

[199] Y. Furukawa, Y. Chikaraishi, N. Ohkouchi, N. O. Ogawa, D. P. Glavin, J. P. Dworkin, C. Abe, and T. Nakamura, "Extraterrestrial ribose and other sugars in primitive meteorites," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 24440–24445, Dec. 2019. Publisher: Proceedings of the National Academy of Sciences.

[200] M. Yadav, R. Kumar, and R. Krishnamurthy, "Chemistry of Abiotic Nucleotide Synthesis," *Chem. Rev.*, vol. 120, pp. 4766–4805, June 2020. Publisher: American Chemical Society.

[201] F. Angiolari, S. Huppert, and R. Spezia, "Quantum versus classical unimolecular fragmentation rate constants and activation energies at finite temperature from direct dynamics simulations," *Phys. Chem. Chem. Phys.*, vol. 24, pp. 29357–29370, Dec. 2022. Publisher: The Royal Society of Chemistry.

[202] C. Schran, F. Brieuc, and D. Marx, "Transferability of machine learning potentials: Protonated water neural network potential applied to the protonated water hexamer," *The Journal of Chemical Physics*, vol. 154, p. 051101, Feb. 2021.

[203] M. Bocus, R. Goeminne, A. Lamaire, M. Cools-Ceuppens, T. Verstraelen, and V. Van Speybroeck, "Nuclear quantum effects on zeolite proton hopping kinetics explored with machine learning potentials and path integral molecular dynamics," *Nat Commun*, vol. 14, p. 1008, Feb. 2023. Number: 1 Publisher: Nature Publishing Group.

[204] G. Adrian and G. Jason, "Accurate Rate Calculations for Hydrogen Atom Abstraction from Methane by Hydroxyl Radical: A Combination of RPMD, Machine Learning Potentials, and Active Learning," *chemrxiv*, May 2023.