Prebiotic chemistry kinetics from stochastic sampling (PROCESS)

Léon Huet

Sorbonne Université

École doctorale 397: Physique et Chimie des Matériaux

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

PRebiOtic Chemistry kinEtics from Stochastic Sampling (PROCESS)

Thèse de Doctorat en Physique

Présentée par Léon Huet

Dirigée par

A. Marco Saitta

Co-encadrée par

Fabio Pietrucci et Rodolphe Vuilleumier

À présenter et soutenir le 20 Septembre 2024

Devant un jury composé de:

Présidente du Jury:	Marie-Laure Bocquet	ENS, Département de Physique - Sorbonne Université
Rapporteurs:	Noel Jakse	SIMAP - Grenoble INP Université Grenoble Alpes
	Ilaria Ciofini	i-CLeHS - Chimie ParisTech PSL
Directeur:	A. Marco Saitta	IMPMC - Sorbonne Université
Co-encadrants:	Fabio Pietrucci	IMPMC - Sorbonne Université
	Rodolphe Vuilleumier	ENS, Département de Chimie - Sorbonne Université
Examinateurs:	Cornelia Meinert	Institut de Chimie de Nice - Université Côte d'Azur
	Mounir Tarek	LPCT - Université de Lorraine

Remerciements

Acknowledgments are written in English and French to be understood by the people to whom they are addressed. Les remerciements sont écrits en partie en anglais et en partie en français pour être compris par les personnes à qui ils sont adressés.

I would like to thank Ilaria Ciofini and Noel Jakse, who accepted to review this thesis. In addition, I thank all the members of the jury who have accepted to judge my work and my defense.

Merci à Marco pour avoir dirigé ma thèse et m'avoir soutenu durant ces trois ans, tout particulièrement dans les moments difficiles. Parcourir puis écrire une thèse est un labyrinthe dans lequel je me serais égaré sans aide.

Merci à Fabio et Rodolphe pour avoir co-encadré ma thèse. Nos discussions ont été les points d'ancrage de ma thèse, des références m'indiquant le chemin à suivre.

A general thanks to the "coffee and lunch" team: Matteo, Sonia, Line, David, Mattia, Flavio, Arthur, Zack, Hadrien, Stefano, and many others... A special thanks to Zack for his exceptional coffee. Thanks also to all the table football team.

En particulier, merci à Arthur, Flavio et Hadrien pour leurs conseils et leur soutien envers la tête de mule que je suis. Un grand merci aussi à Mathieu Moog pour nos longues discussions qui m'ont si efficacement sorti du quotidien de mon stage puis de ma thèse.

Comment oublier l'ensemble de l'orchestre symphonique Alfred Lœwenguth avec qui j'ai pu continuer à jouer pendant la thèse. C'est bien grâce à cela que je n'ai pas perdu la maîtrise de mon instrument de musique. Je les remercie tous chaleureusement pour leur accueil. Les concerts étaient tous formidables.

Je remercie également Laure, Mickaël, Manon, Thibaut, Clément, Florent, Juliette, Lucile et Jean-Vincent pour les longues soirées dans les bars à jeux ou autour d'une table chez l'un d'entre nous ; une chaleureuse assemblée en toutes circonstances. Merci aussi à eux pour ces super vacances tout au long de la thèse.

Merci à Florian et Marina pour leur amitié. Je leur souhaite de bien s'installer à Rennes.

Merci à mon frère Armand et à ma sœur Myrtille pour leur présence à mes côtés durant ces trois ans. Merci à mes parents Isabelle et Franck pour leur soutien inconditionnel pendant toutes mes études supérieures qui ne se sont pas déroulées sans accrocs. Je ne mesure pas ma chance. Merci plus généralement à toute ma famille.

Il est dur de trouver les bons mots quand on en vient à la personne avec qui on n'arrive

pas à imaginer ce qu'il se serait passé sans elle. Il y a des relations qui transforment la vie et donnent la volonté de continuer face à ses peurs. Je remercie Célia pour tout, et c'est bien peu dire.

Pour terminer mes remerciements, j'aimerais remercier mon parrain, Jean-François, qui m'a un jour encouragé à faire de la chimie lors d'un de ses passages à Paris où nous sommes allés à la Cité des Sciences il y a longtemps. Il m'y avait offert un kit pour faire pousser des cristaux.

Contents

Résumé (en Français)

Les origines de la vie sur Terre demeurent l'une des questions de recherche les plus fascinantes, la chimie prébiotique fournissant un aperçu de la manière dont des molécules organiques simples ont pu se former et évoluer avant l'apparition de la biochimie complexe des organismes vivants. Dans ce domaine, les simulations offrent des informations précieuses sur les différents scénarios conduisant à une synthèse réussie ou échouée pour certains composés clés. Cette thèse explore l'étude théorique des réactions de chimie prébiotique en solution aqueuse.

L'étude des réactions chimiques en solution est un domaine complexe et diverse. En utilisant la dynamique moléculaire *ab initio*, nous modélisons avec précision les mouvements des molécules en solution. Cependant, le pas de temps requis pour échantillonner correctement cette évolution est de nettement plus petit que le temps nécessaire pour que les réactions chimiques se produisent. Associé à l'utilisation de la mécanique quantique pour évaluer les forces, cela fait qu'il est impossible d'attendre que jes composé réagisse spontanément, comme ce serait le cas dans des conditions expérimentales.

Pour surmonter ces limitations, des techniques d'échantillonnage avancées telles que la métadynamique, le transition path sampling et l'umbrella sampling sont employées pour explorer de manière approfondie les surfaces d'énergie potentielle et les mécanismes de transition. Ces méthodes permettent l'identification des chemins réactionnels, l'échantillonnage des chemins de transition les plus probables, la création de coordonnées réactionnelles basées sur des données, et le calcul des profils d'énergie libre le long de ces coordonnées.

En complément de ces analyses thermodynamiques, nous étudions la possible application d'équations de dynamique stochastique non-Markovienne pour estimer la cinétique de ses réactions. Cette approche tient compte des effets de mémoire et des corrélations temporelles complexes dans les mouvements moléculaires, offrant des estimations plus rigoureuse des taux de réaction par rapport aux méthodes traditionnelles.

Dans un premier temps, nous appliquons notre protocole de travail à une réaction emblématique S_N2 dans de l'eau explicite. Nous présentons des outils pour mesurer efficacement la qualité des coordonnées réactionnelles potentielles au sommet de la barrière de transition.

Dans un deuxième temps, nous appliquons cette méthode à un mécanisme inédit en 8 étapes pour la synthèse de la glycine, découverte dans notre laboratoire grâce à une simulation de métadynamique. Nous combinons notre cadre initial avec des potentiels machine learning pour générer de nouvelle données, réduisant le coût de la simulation par un facteur 10.

Dans un troisième temps, nous explorons l'utilisation potentielle des modèles non-Markoviens pour l'inférence cinétique dans les réactions chimiques. Nous abordons les défis restants pour l'application agnostique de ces outils et démontrons leur fiabilité dans

l'inférence des taux cinétiques en utilisant des coordonnées réactionnelles heuristiques.

Dans l'ensemble, cette thèse souligne l'intégration des outils computationnels avancés dans la chimie prébiotique. En combinant la dynamique moléculaire *ab initio*, les techniques d'échantillonnage avancées, la dynamique non-Markovienne et les potentiels machine learning, nous fournissons un cadre complet pour explorer et étudier les réactions prébiotiques. Les méthodologies et résultats présentés offrent de nouvelles perspectives sur la première synthèse de molécules organiques simples et ouvrent la voie à de futures recherches sur les origines de la vie et sur l'inférence de la cinétique des réactions chimiques en solution.

Abstract (in English)

The origins of life on Earth remain one of the most fascinating questions in science, with prebiotic chemistry providing critical insights into how simple organic molecules could have formed and evolved, and then leading to the appearance of the complex biochemistry of living organisms. In this domain, atomistic computer simulations offer valuable insights into different scenarios that lead to successful or failed synthesis for certain key compounds. This thesis explores the theoretical study of prebiotic chemical reactions in aqueous environments.

The study of chemical reactions in solution is a complex and multifaceted field. Using *ab initio* molecular dynamics, we accurately model the movements of molecules in solution. However, the typical simulation time step required to adequately sample this evolution is many orders of magnitude smaller than the time required for chemical reactions to occur. Combined with the use of quantum mechanics to evaluate forces, this makes it highly impractical to wait for reactions to occur spontaneously as they would in experimental settings.

To address these limitations, enhanced sampling techniques such as metadynamics, transition path sampling, and umbrella sampling are employed to explore potential energy surfaces and transition mechanisms extensively. These methods enable the identification of reaction pathways, the sampling of reactive regions, the definition and training of data-based reaction coordinates, and the calculation of free energy profiles along these coordinates.

In addition to these thermodynamic analyses, we apply non-Markovian stochastic dynamics equations to estimate reaction kinetics. This approach accounts for memory effects and complex temporal correlations in molecular motions, offering more accurate estimations of reaction rates compared to traditional methods.

In the first step, we present the results of our agnostic workflow applied to an emblematic S_N2 reaction in explicit water. We introduce tools for effectively measuring the quality of potential reaction coordinates at the top of the transition barrier.

In the second step, we apply our method to a novel 8-step mechanism for the synthesis of glycine, discovered in our group through metadynamics exploration. We combine our initial framework with machine learning interatomic potentials for training and sampling, reducing the simulation cost by a factor of 10.

In the third step, we investigate the potential use of non-Markovian models for kinetic inference in chemical reactions. We address some remaining challenges for the agnostic application of these tools and demonstrate their reliability in inferring kinetic rates using heuristic reaction coordinates.

In general, this thesis highlights the integration of advanced computational tools in prebiotic chemistry, marking a significant advancement in both theoretical and practical approaches. By combining ab initio molecular dynamics, enhanced sampling techniques,

non-Markovian stochastic dynamics, and machine learning interatomic potentials, we provide a comprehensive framework for exploring and understanding prebiotic reactions. The methodologies and findings presented here offer new perspectives on the synthesis of complex organic molecules and set the stage for future research on the origins of life and the inference of kinetics for chemical reactions in solution.

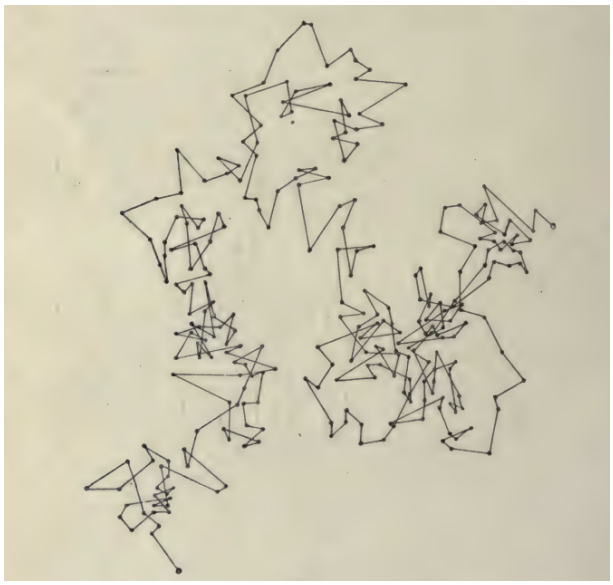


Figure 0.0.1: The Brownian motion as observed in 1916 by Jean Perrin and reported in his book *Atoms*[2]

Introduction

This Introduction is voluntarily simplified to be understood by nonspecialists.

If an adjective can best describe the behavior of matter, it should be "stirred". The matter is agitated. In a gas, a solid, or a liquid, atoms move very fast. The average quadratic speed of an atom at 0° C is 1 km per second, regardless of its environment. This is a reality hard to imagine when one simply looks at a steady cube of ice.

A common way to "observe" this velocity is during rapid depressurization events, such as the famous "can crusher" experiment[1]. The process is as follows: An experimenter boils a small quantity of water in an aluminum can, then, using a pair of tongs, flips it into a container with cold water, submerging the opening. At that exact moment, two events occur in sequence and very quickly. First, the temperature in the can drops rapidly as a result of the cold water and because the aluminum can conducts heat very efficiently. The water vapor inside then condenses into liquid water, creating a low-pressure zone within the can. Instantaneously, the can is crushed by the atmospheric pressure outside, which is no longer balanced by the pressure inside.

To explain the link with the velocity of the atoms, we adopt a nanoscopic point of view. Due to their agitation, the atoms and molecules in the air hit the borders of the can with very high velocities, transmitting this velocity (or more rigorously, their momentum) to the atoms of the can's borders. As there is no longer compensation for this effect from the inside of the can, all borders quickly acquire a significant velocity directed toward the center of the can, causing it to crush. The momentum of this movement is provided entirely by the agitation of the molecules in the air.

Another way to directly observe the velocity of atoms, although more challenging, is

by studying the behavior of a pollen grain on the surface of a pot of hot water. The pollen grain exhibits a movement that appears random. This movement is illustrated in Figure 0.0.1, a figure drawn by Jean Perrin in 1916[2]. The pollen grain seems to bounce in all directions, changing its orientation at very short intervals. This is due to collisions between the pollen grain and the solvent molecules. Both parts of the system exchange momentum, and at this scale, matter is discrete, so this transfer is also discretized and occurs through multiple successive bounces. The observation and quantification of this phenomenon was one of the methods used by Jean Perrin to prove the existence of atoms.

The first question that arises from these observations is why matter is not just a gas. Given the high level of agitation, how can matter remain "condensed" with atoms hitting each other at such high speeds? In solids and liquids, the kinetic energy that results from the movement of the atoms is not sufficient to counteract the electrostatic interactions between atoms and molecules. They attract and link with each other, creating stable structures. Solid and liquid materials are stable for the same reason that an elastic band returns to its original position after being stretched.

The second question is why matter is so complex. With atoms constantly moving, it should remain homogeneous. If black sand and white sand are shaken together, the mixture should turn gray. Thermic agitation has a natural tendency to make the mater mixed. However, atoms are interacting within each other in a really complex manner and this diversity of interaction is what generates certain forms of organization. At 25°C in water, agitation is sufficient to constantly break and form bonds between the molecules, but not sufficient to break the bonds within the molecules at the same rate. This means that in pure water, the vast majority of atoms are organized into H_2O molecules. Additionally, solvated molecules can remain stable for a long time despite the agitation of the solvent, allowing them to react with other compounds and diffuse in the liquid. This is because the covalent bonds that ensure molecular cohesion are more stable than the van der Waals forces or hydrogen bonds between molecules. In terms of energy, covalent bonds are between 100 and 500 $kJ.mol^{-1}$, hydrogen bonds are between 1 to 40 $kJ.mol^{-1}$ and van der Waals forces are of the order of magnitude of $1kJ.mol^{-1}$. At 25°C, the internal kinetic energy is around $3kJ.mol^{-1}$ which is sufficient to regularly break the intermolecular links but not the covalent links.

In this complex system with 3 orders of forces, the diversity of possible phenomena makes it difficult to predict their behavior without a complete simulation.

Although the general principles of complexity emergence in condensed matter are known [3], the diversity of possible phenomena makes it difficult to predict the behavior of these systems. Among the phenomena that are still only coarsely understood is the emergence of life. Life on Earth emerged due to many favorable factors [4]. However, the most important factor, according to the majority of researchers, is the presence of liquid water[5]. In liquid water, the first ingredients of life could have organized themselves to form the first organic compounds.

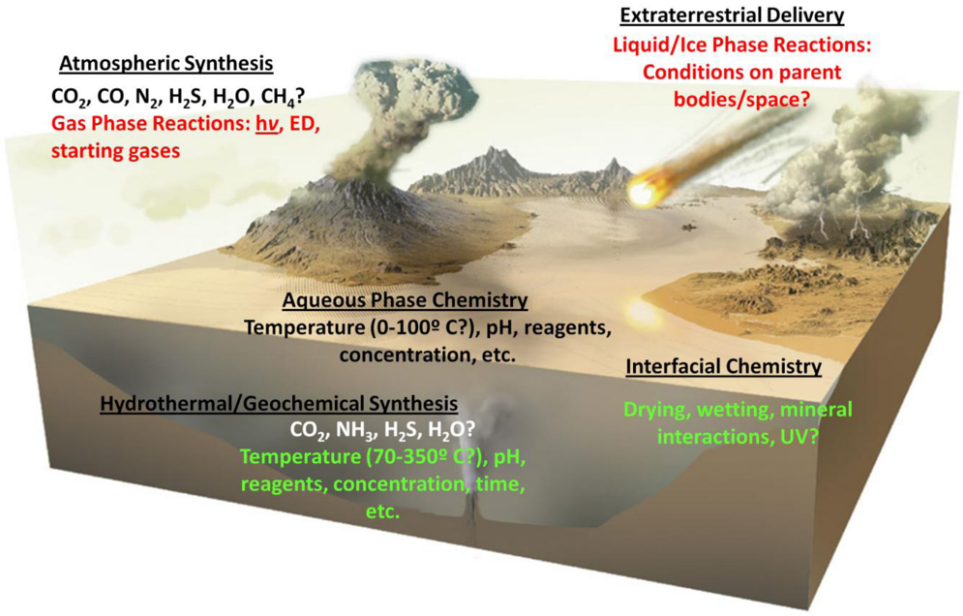


Figure 0.0.2: An artist's representation of different possible scenarios for the synthesis of the first organic compounds on primordial Earth, taken from [6] and adapted from [7]

As presented in Figure 0.0.2, taken from the thesis of Timothée Devergne [6] and adapted from[7], it is currently challenging to determine which of these compounds were synthesized on Earth and which arrived via meteorites, as well as how these processes occurred. However, what is certain is that at some point in the history of life, complexity arose from natural phenomena without the biological machinery present in current living forms. This field of study is known as prebiotic chemistry: the formation of the simplest building blocks of life in an environment where life does not yet exist.

The first person who proposed the natural emergence of life from non-living matter was Charles Darwin with his concept of primordial soup[8]. However, it took more than fifty years before this idea evolved into a subject of experimental investigation. In 1953, Miller and Urey conducted an experiment in which they generated many amino acids and other organic molecules using a setup that contained only inorganic materials: H_2O, CH_4, NH_3 and H_2[9]. This experiment opened the door for research into the synthesis pathways of prebiotic molecules, demonstrating that such synthesis from inorganic matter was possible in a laboratory setting.

More recently, the use of computer simulations has gained significant importance in this field, starting with the reproduction of Miller's work[10]. The advantage of simulations is that they allow the observation of all intermediates in a mechanism, whereas experiments typically only reveal the most stable compounds. Following this initial simulation, numerous periodic studies have been conducted using computational methods[11], [12], [13].

In this paper, we make predictions for chemical reactions in a liquid environment to explore some of the great possibilities for synthesizing these initial building blocks of life.

This requires tools that enable us to model water and the chemical compounds dissolved within it. Additionally, we must be able to model the frenetic thermal agitation of water while simultaneously studying chemical reactions that occur on completely different timescales.

To achieve this complex task, we rely on a precise and transferable method for modeling the behavior of matter: *ab initio* molecular dynamics. This tool is the foundation of this work. It uses the laws of physics (especially quantum physics) to generate simulations that predict the evolution of the studied system over time. From these simulations, further analyses can be performed to make predictions, with the reliability of the results always depending on the accuracy of the way we generate the data.

The first part of this work will introduce the tools necessary for studying chemical reactions in solution, covering both thermodynamics (the energy involved in the studied transformations) and kinetics (the time it takes for these transformations to occur).

The second part will present three completed works directly in their article forms. First, we explore the development of new tools for generating reaction coordinates for chemical reactions in solution. Then, these tools are applied to the study of a new pathway for glycine synthesis under prebiotic conditions, which is an alternative to the commonly accepted Strecker pathway[14]. The final chapter will introduce a new framework designed to efficiently separate the behavior of a reaction coordinate from the other degrees of freedom in the simulation using a non-Markovian model of stochastic equations. This stochastic model can be used to obtain important information about the dynamics of the variable and also to calculate the kinetics efficiently.

Part I

Methods: molecular dynamics and stochastic models for chemical reactions in solutions

Chapter 1

Molecular dynamics: from classical to quantum

Introduction to the idea of *ab initio* and molecular dynamics calculations

Theoretical chemistry is a vast and growing field. It encompasses many concepts, some of which are very fundamental and close to fundamental physics, while others are very empirical, aiming to directly explain the complexity of observations "made on the bench". We could roughly categorize these two branches as follows: the branch of atomistic modeling and the branch of thermochemistry.

Today, the theoretical bridge between these two complementary branches is still not completely established. The issue is mainly due to the entropy[15, 16, 17]. Today, it is permissible to hope that the atomistic branch, the more fundamental one, will fully catch up with empirical observations in the case of chemistry. However, thermochemistry is still predominantly used in laboratories despite its limitations. This Introduction will touch on the challenges ahead to cross this point.

Water

A simple glass of 20cl pure water contains about $3 \cdot 10^{25}$ atoms. Modeling such a system atom by atom is unreachable. However, one can see that the water is very well simulated in our laboratories. Every day, engineers and researchers perform very precise fluid mechanics simulations and thus manage to optimize very complex objects, such as the hydrodynamic shape of a boat's hull. Clearly, these two facts are in contradiction. The solution, however, is relatively simple: we do not model water itself, just an idea of water.

When people model what they call "water" on a large scale, they replicate its emergent properties: its density, its temperature, its pH, its chemical composition, its heat capacity, its compressibility, its viscosity, *, etc.*. We then use a continuous model, for example,

fluid dynamics[18]. No large-scale model reproduces water itself in all its complexity. For example, one can imagine adding another material to the system that will react with water and change its properties; no macroscopic system could make real predictions about what would happen without manually adding new properties specific to this addition.

Empirical to *ab initio*

Macroscopic models are "empirical"; they accumulate a lot of data derived from the observation and measurement of macroscopic phenomena. They are also poorly transferable, as they are very specific to certain types of problem and cannot solve anything else without someone manually modifying the type of calculation performed. However, these models are those that allow us to predict whether planes fly, ships float, and phones call.

In contrast, the *ab initio* models (in Latin: from the beginning) are more ambitious. The objective here is to use as few external data as possible and to calculate everything else from a theoretical base considered reliable. Here, the application domain of the model is as vast as the validity domain of the underlying physical theory. In the case of chemistry, this mother theory is quantum mechanics, the only theoretical level capable of explaining the behavior of matter, and notably the covalent bond, the backbone of chemistry.

Purely *ab initio* calculations are rare. In all the calculations we will present here, contrary to what this word implies, many premises come from other sources of knowledge: atomistic, chemistry, and classical physics. In the field, many researchers still rely on the "intuition of the chemist", at least partially. These premises, based on observation and not purely theoretical arguments, are difficult to detect. They are found in our very first positions of atoms, in the variables we decide to observe, in the way we interpret some results *, etc.* In general, they are found mainly when it comes to making a decision. When these *errors* are understood and accepted by the research community, it is not a major problem. Often, they are necessary, extensively verified, or justified by solid arguments absent from the original theory. *Ab initio* is therefore more a guideline whose deviations are the sensitive points of our models that we have to take care of.

Complexity

The complexity of quantum mechanics can be illustrated in several ways. Here, we will focus on two fundamental aspects.

For a system with more than two particles treated by quantum mechanics, exactly solving the Schrödinger equation is typically infeasible[19]. As a result, one must make approximations, which means that one must stress some accuracy of the underlying theory even before beginning the calculation.

Furthermore, the complexity of the calculation is not proportional to the number of electrons (N_e) in the system; this complexity evolves as 2^{N_e} in the most expensive cases. In the case of this thesis, it will be more like ${N_e}^3$, which is much more reasonable but still

limits the calculation to the order of a thousand atoms.

This limitation proves problematic not only because the expected system size is unmet but also because at the scale of a thousand atoms, macroscopic quantities, such as pH, density, heat capacity, and molecular concentration, are inadequately defined. These macroscopic measurements represent averages that exhibit greater stability when derived from 1 mol (10^{23}) of atoms rather than merely a thousand. For example, when modeling 100 water molecules, a sudden transformation of one molecule into an oxonium-chloride ion pair results in an abrupt pH transition from 7 (neutral) to 0.15 (very acidic), without intermediate values. There exists a method to address these discontinuities: if an oxonium ion is observed to be present "1% of the time" in the modeling box, this can be construed as an "apparent" intermediate concentration on a macroscopic scale, yielding an intermediate pH of 2.2. This approach substitutes the necessity for a large atomic count with the requirement for extensive temporal observations, underscoring the rationale for conducting dynamics.

1.1 Newton's law integration and temperature control

Molecular dynamics consists of the study of the behavior of a certain number of atoms (or molecules) over time. The positions of these atoms can be represented by the N 3D vectors $\mathbf{R_i}$, with N being the number of atoms. In fact, it is the positions of the nuclei: the positions of the electrons are not considered yet. These positions can be concatenated into a vector of dimensions $3N$: $\mathbf{R}$, which encloses the entire geometry of the system. The evolution of this vector is calculated by integrating the equation of motion, which is the second law of Newton.

For a nucleus i among the N ones:

$$m_i \ddot{\mathbf{R}}_\mathbf{i} = f_i(\mathbf{R}) \tag{1.1.1}$$

where m_i designates the mass of the atom i, and $f_i(\mathbf{R})$ is the force applied on the atom i for a geometry $\mathbf{R}$.

One can define M a (3N,3N) diagonal matrix containing the mass of each of the N atoms, three times in a row for the 3 dimensions:

$$M = \begin{pmatrix} m_1 & & & & & & & & \\ & m_1 & & & & & & & \\ & & m_1 & & & & & & \\ & & & m_2 & & & & & \\ & & & & \ddots & & & & \\ & & & & & m_N & & & \\ & & & & & & m_N & & \\ & & & & & & & \ddots & \end{pmatrix} \tag{1.1.2}$$

Using M, we can redefine our initial Newton equation with all dimensions of the system in a single matrix equation:

$$\ddot{\mathbf{R}} = M^{-1}\mathbf{F}(\mathbf{R}) \tag{1.1.3}$$

where F designates the full force vector of dimensions $3N$. We can also define the momentum vector, $\mathbf{P}$, which will be used in the latter.

$$\mathbf{P} = M\dot{\mathbf{R}} \tag{1.1.4}$$

This notation with M the mass matrix and $\mathbf{R}$ the configuration vector will help us in the development of dynamics equations at the end of this part.

Equation 1.1.3 is continuous and it is not possible to calculate any analytical solution for it. The discretization of the integration is the only way to proceed. In this condition, $\mathbf{R}$ is recalculated every Δt time step, causing a certain amount of error at each step. There are many ways to discretize the equation of Newton, but the most commonly used is the Verlet method.

1.1.1 Verlet Algorithm

The Verlet algorithm is based on the following equations [20]:

$$\mathbf{R}(t-\Delta t) = \mathbf{R}(t) - \dot{\mathbf{R}}(t)\Delta t + \frac{\ddot{\mathbf{R}}(t)}{2}\Delta t^2 - \frac{\dddot{\mathbf{R}}(t)}{6}\Delta t^3 + \mathcal{O}(\Delta t^4) \tag{1.1.5}$$

$$\mathbf{R}(t+\Delta t) = \mathbf{R}(t) + \dot{\mathbf{R}}(t)\Delta t + \frac{\ddot{\mathbf{R}}(t)}{2}\Delta t^2 + \frac{\dddot{\mathbf{R}}(t)}{6}\Delta t^3 + \mathcal{O}(\Delta t^4) \tag{1.1.6}$$

$$\mathbf{R}(t+\Delta t) = 2\mathbf{R}(t) - \mathbf{R}(t-\Delta t) + M^{-1}\mathbf{F}(\mathbf{R})(\Delta t)^2 + \mathcal{O}(\Delta t^4) \tag{1.1.7}$$

Equation Eq 1.1.7 is simply the sum of Eq 1.1.5 and Eq 1.1.6. This trick of basing our integration on the two preceding positions $R(t)$ and $R(t-\Delta t)$ to calculate $R(t+\Delta t)$ allows us to diminish the integrating error to $\mathcal{O}(\Delta t^4)$, without the need of a more expensive calculation than $F(\mathbf{R}(t))$.

The velocity must be calculated *a posteriori* using the difference between Eq 1.1.5 and Eq 1.1.6:

$$\dot{\mathbf{R}}(t) = \frac{\mathbf{R}(t+\Delta t) - \mathbf{R}(t-\Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^2) \tag{1.1.8}$$

Δt has to be chosen carefully. The cost of the calculation over a fixed time interval t_0 evolves linearly with $1/\Delta t$. By halving the time steps, the "walk" on the interval becomes slower, making it more computationally expensive. However, using a too large time step can also create problems, especially for phenomena with high velocities in the system, such as high-frequency oscillations. For them sufficiently small time steps are required for accurate modeling. Employing excessively large integration steps can lead to nonphysical phenomena due to integration errors.

The ensemble of all geometries encountered during the process is called a trajectory and can be denoted as $(\mathbf{R_n})$. We can also define $(\mathbf{P_n})$ as the corresponding momentum trajectory using the following equation:

$$\forall n, \mathbf{P_n} = M\dot{\mathbf{R}}_n \tag{1.1.9}$$

1.1.2 Periodic boundary conditions

To model a liquid phase, we use a cubic box of the lattice a and apply periodic boundary conditions (Periodic Boundary Conditions (PBC)). This approach allows us to simulate a liquid phase with fewer than hundreds of molecules. The periodic boundary conditions are implemented using the following equation:

$$\mathbf{R}(t) = \mathbf{R}(t) - a \cdot \text{floor}\left(\frac{\mathbf{R}(t)}{a}\right) \tag{1.1.10}$$

where floor is a function that truncate each number inside the $\frac{\mathbf{R}(t)}{a}$ vector to the closest lower integer.

With a fixed lattice box, the volume of the system remains constant: $V = a^3$. We can adjust a with respect to the quantity of compounds in the box to obtain a realistic density: $\rho = \frac{m_{tot}}{a^3} \approx 1kg.L^{-1}$.

To improve accuracy and better reflect the manipulation of the bench, we could introduce variations in the lattice of the box, *, that is,*, add a barostat. This is because experimentalists often work at constant pressure rather than at constant volume. However, for chemical reactions in high-dilution contexts, we consider these volume variations to be negligible.

1.1.3 Thermostat: Nosé-Hoover, single and chain

The Verlet algorithm, as derived from the second law of Newton, models a completely isolated system, without any exchange with the outside. This is rather unrealistic in experimental chemistry conditions, where the temperature of the system is often controlled. To model this, we need to intervene in the Verlet propagation equation by adding an external friction term that will correct the velocities of the system toward a target temperature: T_0.

This friction term can take many forms. In this thesis, we used the Nosé-Hoover chain thermostat primarily[21]. To explain it step by step, we first explain with the single Nosé-Hoover case[22, 23].

The instantaneous temperature of a system, T, can be directly linked to $\dot{\mathbf{R}}$, the velocity of the particles in the system.

$$\begin{aligned} \frac{1}{2}N_d k_B T &= \frac{1}{2}(\dot{\mathbf{R}}^{\mathrm{T}} M \dot{\mathbf{R}}) \\ &= \frac{1}{2}(\mathbf{P}^{\mathrm{T}} M^{-1} \mathbf{P}) \end{aligned} \tag{1.1.11}$$

$$N_d = (3N - N_c) \tag{1.1.12}$$

Where N_d is the number of degrees of freedom and N_c is the number of constraints (blocked translations and rotations).

The idea behind the Nosé-Hoover thermostat is to add an external term to the evolution equation that interacts by friction with all other atoms in the system to maintain a fixed temperature T_0. This external fictive variable, η, has the dimension of the inverse of time and is associated with a pseudo-mass Q, which determines the thermostat's inertia.

Using equation 1.1.11, we can define the single Nosé-Hoover thermostat by modifying Newton's equation 1.1.3 as follows:

$$\begin{cases} \ddot{\mathbf{R}} = M^{-1}F - \eta\dot{\mathbf{R}} \\ \dot{\eta} = \dfrac{1}{Q}N_d k_B(T - T_0) \\ \quad = \dfrac{1}{Q}\left(\mathbf{P}^{\mathrm{T}}M^{-1}\mathbf{P} - N_d k_B T_0\right) \end{cases} \tag{1.1.13}$$

The mass Q can be fixed using a more intuitive parameter, the characteristic time scale τ:

$$Q = N_d k_B T_0 \tau^2 \tag{1.1.14}$$

However, it has been shown that for small systems, the Nose-Hoover thermostat does not guarantee an efficient sampling of the velocities and positions (the notion of ergodicity involved here will be defined in the next subsection). This comes from the η fluctuations that are not sufficiently driven toward a Gaussian distribution as should be. To overcome this problem, we can protect the undriven variable under layers of thermostats. This is the idea behind the Nose-Hoover Chain (NHC) [21]. The following equation system presents an n-layered thermostat:

$$\begin{cases} \ddot{\mathbf{R}} = M^{-1}F - \eta_1\dot{\mathbf{R}} \\ \dot{\eta}_1 = \dfrac{1}{Q_1}\left(\mathbf{P}^{\mathrm{T}}M^{-1}\mathbf{P} - N_d k_B T_0\right) - \eta_1\eta_2 \\ \dot{\eta}_j = \dfrac{1}{Q_j}(Q_{j-1}\eta_{j-1}^2 - k_B T_0) - \eta_j\eta_{j+1} \qquad ; \quad (2 \leq j < n) \\ \eta_n = \dfrac{1}{Q_n}(Q_{n-1}\eta_{n-1}^2 - k_B T_0) \end{cases} \tag{1.1.15}$$

$$\begin{cases} Q_1 = N_d k_B T_0 \tau^2 \\ Q_j = k_B T_0 \tau^2 \qquad ; \quad (2 \leq j) \end{cases} \tag{1.1.16}$$

Most of the time, n is not greater than 5. In the case of this work, we will systematically use a 3-layer thermostat. This method has been shown to improve the quality of the velocity distribution, especially in the case of small systems with sharp forces[21]. The determinism of the simple Nose-Hoover version is kept, along with the possibility to reverse time in these equations.

1.2 Introduction to Statistical Mechanics

Now that we have a method to sample our system, we need to establish a way to connect our findings with macroscopic observations. We took a step in this direction when we defined the Nose-Hoover thermostat and the instantaneous temperature (see subsection 1.1.3). All typical physical quantities that can be measured in the laboratory are defined over an enormous number of particles. These quantities are called *observables*. Among them are temperature, pressure, pH, density, *, etc.*. Since a large number of degrees of freedom are involved in each of these quantities, the best way to link them with nanoscale results is to use statistical tools, following the approach of Boltzmann[24].

1.2.1 Observable: from micro to macro

Let us define a macrosystem (typically with a size larger than the micrometer scale) called $\mathcal{S}$. It is made up of a large ensemble of microsystems: $\mathcal{S} = \{s\}$. All microsystems s are considered equivalent. Each microsystem has an instantaneous state, which is fully defined by $(\mathbf{R}, \mathbf{P})$, the position and momentum of the particles inside it (see section 1.1). The ensemble of all possible values of the couple $(\mathbf{R}, \mathbf{P})$ is called the phase space. At every instant, the probability density to find one system s in the exact state $(\mathbf{R}, \mathbf{P})$ is designated as $\rho(\mathbf{R}, \mathbf{P})$. An observable of the microsystem is a function of the position and velocities of s: $O(\mathbf{R}, \mathbf{P})$. A macroscopic intensive measurement in $\mathcal{S}$ associated with this observable will return a value $\mathcal{O}$, which is typically the average value of $O(\mathbf{R}, \mathbf{P})$ across all microsystems $\{s\}$:

$$\mathcal{O} = \iint \rho(\mathbf{R}, \mathbf{P}) O(\mathbf{R}, \mathbf{P}) d\mathbf{R} d\mathbf{P} \tag{1.2.1}$$

Using this, we only need to represent one geometry of the full macroscopic system to obtain the instantaneous probability density $\rho(\mathbf{R}, \mathbf{P})$ and calculate $\mathcal{O}$. However, even a single geometry of the macrosystem is unaffordable due to its size (see introduction of chapter 1). To overcome this, we must limit our observation to the equilibrium probability density $\rho_{eq}(\mathbf{R}, \mathbf{P})$, because this particular case can be estimated using the ergodic hypothesis.

This hypothesis is central in statistical mechanics. It relates the infinite-time average of an observable in one microsystem to the equilibrium probability density of all $\{s\}$.

$$\lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau O(t)\, dt = \iint \rho_{eq}(\mathbf{R}, \mathbf{P})\, O(\mathbf{R}, \mathbf{P})\, d\mathbf{R}\, d\mathbf{P} = \mathcal{O}_{eq} \tag{1.2.2}$$

This allows us to study $\mathcal{S}$ by modeling only one s. However, the properties of s affect the accuracy with which $\mathcal{S}$ represents a realistic system. The next sections are dedicated to explaining how the tools we use to model s (see chapter 1) impact our results.

1.2.2 Extended Hamiltonian and (N,V,T) ensemble

The system is thermostated with the Nose-Hoover Chain (NHC) (see subsection 1.1.3). The total energy of the system can be described by an extended Hamiltonian [21]:

$$\mathcal{H}' = \mathcal{H}(\mathbf{R}, \mathbf{P}) + \sum_{j=1}^{n} \frac{1}{2} Q_j \eta_j^2 + N_d k_B T_0 \eta_1 + k_B T_0 \sum_{j=2}^{n} \eta_j \tag{1.2.3}$$

Here, $\mathcal{H}$ is the Hamiltonian of the system prior to the addition of the thermostat. It is composed of a potential part, V, detailed in the next section, and a kinetic part:

$$\mathcal{H}(\mathbf{R}, \mathbf{P}) = V(\mathbf{R}) + \frac{1}{2} \mathbf{P}^{\mathrm{T}} M^{-1} \mathbf{P} \tag{1.2.4}$$

The equations of motion of the NHC system (see Eq 1.1.15) are derived from $\mathcal{H}'$. In his 1984 paper, Nosé proved that if the behavior of $(\mathbf{R}, \mathbf{P}, \{\eta_j\})$ is microcanonical (*i.e.*, the extended Hamiltonian is conserved over time), the phase space $(\mathbf{R}, \mathbf{P})$ follows a canonical distribution at equilibrium [23]. This result has been generalized for the NHC.

Using this, our system can be studied in the (N, V, T) ensemble formalism. We can define $\rho_{eq}(\mathbf{R}, \mathbf{P})$ as follows:

$$\rho_{eq}(\mathbf{R}, \mathbf{P}) = \frac{1}{Z} e^{-\beta H(\mathbf{R}, \mathbf{P})} \tag{1.2.5}$$

where $\beta = \frac{1}{k_B T_0}$, and where Z is the partition function:

$$Z = \iint e^{-\beta H(\mathbf{R}, \mathbf{P})} d\mathbf{R} d\mathbf{P} \tag{1.2.6}$$

The positions and momenta components of the canonical distribution can be separated:

$$\begin{cases} \rho_{eq}(\mathbf{R}) = \dfrac{1}{Z_{\mathbf{R}}} e^{-\beta V(\mathbf{R})} \\ \rho_{eq}(\mathbf{P}) = \dfrac{1}{Z_{\mathbf{P}}} e^{-\frac{1}{2}\beta \mathbf{P}^{\mathrm{T}} M^{-1} \mathbf{P}} \\ Z = \displaystyle\int e^{-\beta V(\mathbf{R})} d\mathbf{R} \int e^{-\frac{1}{2}\beta \mathbf{P}^{\mathrm{T}} M^{-1} \mathbf{P}} d\mathbf{P} = Z_{\mathbf{R}} Z_{\mathbf{P}} \end{cases} \tag{1.2.7}$$

The definition of $\rho_{eq}(\mathbf{R})$ leads us to the marginalized free energy (detailed in the next subsection), while the definition of $\rho_{eq}(\mathbf{P})$ leads us to the Maxwell–Boltzmann distribution of the velocities:

$$\rho_{eq}(v_{i_\alpha}) = \sqrt{\frac{\beta m_i}{2\pi}} e^{-\frac{1}{2}\beta m_i v_{i_\alpha}^2} \quad ; \quad \forall i, \quad \alpha \in \{x, y, z\} \tag{1.2.8}$$

This is the distribution used to initialize velocities in all of our simulations, ensuring a realistic initialization with respect to T_0. This distribution is independent of how the forces are calculated.

1.2.3 Collective variables and free energy

In this manuscript, a Collective Variable (CV) is a differentiable function of the atomic positions: $cv(\mathbf{R})$, where $\mathbf{R}$ is an element of the position part of the phase space: Γ, named the configuration space. Typically, Collective Variable (CV) can be a scalar value or a vector. The CVs are usually described by fewer dimensions than the configuration space which is composed of 3N dimensions. That is why they are often qualified as projectors. This appellation is rather wrong because CVs can be much larger in dimensions than the phase space. The interatomic distances are a typical family of CV and the vector composed of all the interatomic distances which is also a CV by definition is described by $N(N-1)/2$ dimensions which is much larger than $3N$ for large N.

The probability distribution of a CV at equilibrium can be calculated using the phase space position distribution and estimated based on the ergodic hypothesis:

$$\begin{aligned} \rho_{eq}(cv(\mathbf{R})) &= \int \rho_{eq}(\mathbf{R}')\delta(cv(\mathbf{R}) - cv(\mathbf{R}'))d\mathbf{R}' \\ &= \lim_{\tau\to\infty} \frac{1}{\tau}\int_0^\tau \delta(cv - cv(t))d\tau \end{aligned} \tag{1.2.9}$$

In realistic cases, once we have reached apparent stationarity of the sampling, $\rho_{eq}(cv)$ can be estimated based on the histogram of CV.

By analogy with the phase space distribution function, we can define $F(cv)$, the marginalized free energy, also named the free energy profile of the CV:

$$\rho_{eq}(cv) = \frac{1}{Z_{cv}} e^{-\beta F(cv)} \tag{1.2.10}$$

$$F(cv) = A - \frac{1}{\beta} ln(\rho_{eq}(cv)) \tag{1.2.11}$$

F is often defined up to a constant term, as the full partition function of the CV is unknown. Usually, the most probable value of the CV is set to 0 in free energy.

1.3 Quantum mechanics

Now that we have introduced how to sample the dynamics of a system and how we link this to macroscopic measurements, we need to explain how we calculate the forces, which is the central pillar of our work. To explain the sometimes counterintuitive behavior of chemical systems, we must use quantum mechanics, the underlying theory that describes, among others, the behavior of covalent bonds, which ensure the stability of molecules.

However, obtaining an exact solution is infeasible for systems with more than one electron and one nucleus. This necessity leads to a variety of approximations that enable us to compute a solution, scarifying a part of the accuracy. The following sections will describe various approaches, starting from the exact Schrödinger equation and progressively

incorporating more approximations to develop a practical method. Ultimately, these approximations allow us to achieve a sufficiently accurate force estimation to perform *Ab Initio* Molecular Dynamics (AIMD).

1.3.1 Schrödinger equation of a many-body system

Quantum mechanics is based on the system many-body wavefunction, Ψ. This wavefunction depends on the positions of the nuclei $\mathbf{R}$, but also on the positions of the N_e electrons of the system, represented by a $3N_e$ dimension vector $\mathbf{r_e}$, analogous to $\mathbf{R}$. Since the calculation of forces in the Schrödinger equation relies on distances between particles, the electron position vectors $\{\mathbf{r_1}, ..., \mathbf{r_{N_e}}\}$ and the equivalent for nuclei $\{\mathbf{R_1}, ..., \mathbf{R_N}\}$ are also used in the following equations. For the first time, we also use the atomic numbers $\{Z_1, ..., Z_N\}$.

The equation that allows us to determine the time evolution of Ψ and calculate the energies and forces of the system is the time-dependent Schrödinger equation:

$$i\hbar\frac{\partial\Psi(\mathbf{r_e}, \mathbf{R})}{\partial t} = \mathcal{H}_\mathcal{Q}\Psi(\mathbf{r_e}, \mathbf{R}) \tag{1.3.1}$$

Here $\mathcal{H}_\mathcal{Q}$ is the quantum Hamiltonian, to be considered different from the classical $\mathcal{H}$ (see section 1.2). In this case, $\mathcal{H}_\mathcal{Q}$ can be separated into five terms.

$$\mathcal{H}_\mathcal{Q} = T_N + V_{N,N} + V_{N,e} + T_e + V_{e,e} \tag{1.3.2}$$

Two of these terms are associated with the nuclei: T_N, the kinetic energy of the nuclei, and $V_{N,N}$, the potential arising from the Coulombic interaction between the nuclei.

$$\begin{cases} T_N = -\dfrac{\hbar^2}{2}\displaystyle\sum_{i=1}^{N}\frac{\nabla^2_{\mathbf{R_i}}}{m_i} \\ V_{N,N} = \displaystyle\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j>i}}^{N}\frac{Z_i Z_j e^2}{|\mathbf{R_i} - \mathbf{R_j}|} \end{cases} \tag{1.3.3}$$

Two of these terms are associated in the same way but with electrons: T_e and $V_{e,e}$.

$$\begin{cases} T_e = -\dfrac{\hbar^2}{2m_e}\displaystyle\sum_{i=1}^{N_e}\nabla^2_{\mathbf{r_i}} \\ V_{e,e} = \displaystyle\sum_{i=1}^{N_e}\sum_{\substack{j=1 \\ j>i}}^{N_e}\frac{e^2}{|\mathbf{r_i} - \mathbf{r_j}|} \end{cases} \tag{1.3.4}$$

The remaining term is the potential of the cross-coulombic interaction between nuclei and electrons:

$$V_{N,e} = -\sum_{i=1}^{N}\sum_{j=1}^{N_e}\frac{Z_i e^2}{|\mathbf{R_i} - \mathbf{r_j}|} \tag{1.3.5}$$

As we target a stationary solution of the system, we can simplify the initial time-dependent Schrödinger equation (see Eq 1.3.1) to a time-invariant one:

$$\mathcal{H}_{\mathcal{Q}}\Psi(\mathbf{r_e}, \mathbf{R}) = E\Psi(\mathbf{r_e}, \mathbf{R}) \tag{1.3.6}$$

Here, E designates the energy associated with the state represented by $\Psi(\mathbf{r_e}, \mathbf{R})$. E is an eigenvalue of this equation, and if $\Psi(\mathbf{r_e}, \mathbf{R})$ is a solution of the above equation, then it is an eigenvector. The lowest possible eigenvalue E_0 is associated with the ground-state solution of the equation: Ψ_0.

The only systems for which this equation can be analytically solved are the hydrogenoids, composed of 1 electron and 1 nucleus. As it is not possible to make any chemistry with it, we are going to start the descent toward reachable but approximated equations.

The first approximation we can make is the Born-Oppenheimer one[25, 26]. The idea is to consider that the movement of the electrons is much faster than the movement of nuclei by several orders of magnitude, *, i.e.*, from the point of view of the electrons, nuclei are static. This implies that we can separate their behavior and treat them separately. Here, we consider that nuclei have a classical behavior, and electrons have a quantum one, in the static "external" potential generated by the nuclei. We have now decoupled these two calculation parts: nuclei movements are treated classically with Newton's equation and the Verlet algorithm, and electrons are treated separately thanks to the electronic Schrödinger equation; this is the Born-Oppenheimer Molecular Dynamics (BOMD). Now, the positions of the nuclei are not a variable of the wavefunction. They are an external parameter of an electronic Hamiltonian. The terms of this new Hamiltonian that only involve nuclei (see Eq 1.3.3) can be discarded as constants. This approximation is possible because of the three orders of magnitude that separate electron and nucleon masses.

$$\mathcal{H}_{\mathcal{Q}e\{\mathbf{R}\}} = T_e + V_{e,e} + V_{e,N} \tag{1.3.7}$$

This approximation allows us to determine the solutions of the Schrödinger equation for systems larger than the hydrogenoids under the condition that it only contains one electron. In the present case, this also neglects the quantum behavior of the nuclei, which is not true in realistic systems, especially for hydrogen, which can undergo proton hopping through the tunneling effect [27, 28, 29, 30]. To partially overcome this problem and also to increase the numerical stability of the Verlet algorithm (see subsection 1.1.1), theoreticians often replace hydrogen atoms with deuterium atoms, which are two times more massive. Within the Born-Oppenheimer scheme, the electronic Hamiltonian, *i.e.*, the resulting potential, is not impacted by this change; it only modifies the dynamics of the system. Again, in realistic systems, we know that this is not fully true as the enthalpy

of formation of normal water in the gaseous phase (H_2O) is not the same as for heavy water (D_2O)[31].

From this point the potential initially stated as V in the Hamiltonian equation of nuclei movements (see Eq 1.2.4) can be calculated using the Hellman-Feynman theorem[32].

To build on this electronic Hamiltonian, we have to make approximations of the behavior of the electrons.

1.3.2 Hartree-Fock approximation

Initially, the electronic wave function can be any function of $\mathbf{r_e}$. The next approximation involves restricting this ensemble of functions on the basis of an external assumption.

One such assumption is that the complete wave function of the system can be decomposed into N_e single-electron orbitals, which are multiplied together to form the full wave function. This approach was proposed by Hartree in 1928 [33].

$$\Psi(\mathbf{r_e}) = \prod_{i=1}^{N_e} \Phi_i(\mathbf{r_i}) \tag{1.3.8}$$

This decomposition is an ansatz, *i.e.*, an educated hypothesis, that simplifies the complexity of the problem, making it solvable.

$\Phi(r_i)$ are called spin orbitals. Here, we consider that the spin of the electron is included in the electron coordinate r_i. This decomposition of the wavefunction leads to monoelectronic Hamiltonians:

$$\mathrm{h}_{e\{\mathbf{R},\mathbf{r_{k\neq i}},\Phi_{k\neq i}\}}\Phi_i(\mathbf{r_i}) = \epsilon_i \Phi_i(\mathbf{r_i}) \tag{1.3.9}$$

$$\mathrm{h}_{e\{\mathbf{R},\mathbf{r_{k\neq i}},\Phi_{k\neq i}\}} = -\frac{\hbar^2}{2m_e}\nabla^2 - \sum_{j=1}^{N} \frac{Z_j e^2}{|\mathbf{r_i} - \mathbf{R_j}|} + e^2 \sum_{\substack{k=1 \\ k\neq i}}^{N_e} \int \frac{|\Phi_k(\mathbf{r'})|^2 d\mathbf{r'}}{|\mathbf{r_i} - \mathbf{r_k}|} \tag{1.3.10}$$

This time, the single-particle equations 1.3.9 are coupled. This implies that they have to be treated simultaneously. To do that, we use a Self-Consistent Field (SCF). A first (generally atomic) guess is proposed for the single-electron wave-functions $\{\Phi_i\}_0$, then a new set of wavefunctions $\{\Phi_i\}_1$ is recalculated using the equation Eq 1.3.9 but with $\{\Phi_i\}_0$ as initial parameters. Subsequently, the new wavefunctions $\{\Phi_i\}_1$ replace $\{\Phi_i\}_0$ to determine $\{\Phi_i\}_2$. This loop continues until we reach a converged set of single-electron wave-functions (up to a threshold on resulting energies and forces), *i.e.*, a stable solution of equation 1.3.9 for every Φ_i.

Slater [34] and Fock identified that the solutions to equation 1.3.9 were not antisymmetric with respect to the exchange of two electrons, which contradicts the fermionic properties of electrons. To overcome this problem, they proposed to use a mathematical tool that exhibits this antisymmetric behavior upon permutations: the matrix determinant. Here is the so-called 'Slater determinant', our new wave function trial form:

$$\Psi(\mathbf{r_e}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \Phi_1(\mathbf{r_1}) & \Phi_2(\mathbf{r_1}) & \Phi_3(\mathbf{r_1}) & \dots & \Phi_{N_e}(\mathbf{r_1}) \\ \Phi_1(\mathbf{r_2}) & \Phi_2(\mathbf{r_2}) & \Phi_3(\mathbf{r_2}) & \dots & \Phi_{N_e}(\mathbf{r_2}) \\ & & \dots\dots\dots\dots\dots & & \\ \Phi_1(\mathbf{r_{N_e}}) & \Phi_2(\mathbf{r_{N_e}}) & \Phi_3(\mathbf{r_{N_e}}) & \dots & \Phi_{N_e}(\mathbf{r_{N_e}}) \end{vmatrix} \tag{1.3.11}$$

Now that we have determined the type of system we are using, we need to find the ground state, which is the common electronic state of molecules in the absence of light absorption and emission. To do this, we rely on the variational principle[35].

Variational principle

The ground-state energy is determined using the following minimization process:

$$E_0 = \min_{\Psi} \frac{\langle \Psi | \mathcal{H}_e | \Psi \rangle}{\langle \Psi | \Psi \rangle} \tag{1.3.12}$$

In our case, Ψ is a Slater determinant, *i.e.*, a determinant function of Φ_i. To achieve minimization, as in the case of simple products, we can decompose the problem into coupled single-electron eigenvalue equations:

$$h_{HFi}\Phi_i(\mathbf{r_i}) = \epsilon_i \Phi_i \mathbf{r_i}) \tag{1.3.13}$$

Where h_{HF} is defined as:

$$h_{HFi} = -\frac{\hbar^2}{2m_e}\nabla^2 - \sum_{j=1}^{N} \frac{Z_j e^2}{|\mathbf{r_i} - \mathbf{R_j}|} + V_{HF}(\mathbf{r_i}) \tag{1.3.14}$$

V_{HF} is the Hartree-Fock potential energy resulting from the interaction of the electron i with the mean field created by all other electrons. Like the full Hamiltonian, it can be decomposed into two terms, each of which is a sum over all the other electrons:

$$V_{HF}(\mathbf{r_i}) = \sum_{k=1}^{N_e} J_k(\mathbf{r_i}) - \sum_{k=1}^{N_e} K_k(\mathbf{r_i}) \tag{1.3.15}$$

The term $J_k(\mathbf{r_i})$ represents the Coulomb operator, which accounts for the Coulomb repulsion experienced by the electron i due to the rest of the electrons. It is defined as:

$$J_k(\mathbf{r_i}) = e^2 \int \frac{|\Phi_k(\mathbf{r'})|^2 \, d\mathbf{r'}}{|\mathbf{r_i} - \mathbf{r_k}|} \tag{1.3.16}$$

The other term arises purely from quantum mechanics and is known as the exchange term. It is given by:

$$K_k(\mathbf{r_i})\Phi_i(\mathbf{r_i}) = e^2 \int \Phi_k(\mathbf{r'}) \frac{1}{|\mathbf{r_i} - \mathbf{r'}} \Phi_i(\mathbf{r'})\Phi_k(\mathbf{r_i}) \, d\mathbf{r'} \tag{1.3.17}$$

The eigenvalue equations are then solved self-consistently.

Now that we have defined the Hartree-Fock exchange term and introduced the concepts of ansatz, variational principle, and SCF, we can proceed to discuss the density functional.

1.3.3 Density Functional Theory

For a long period in the history of quantum mechanics, the density of the electron cloud of a molecule (*i.e.*, the electronic density) was considered a potential functional for determining the energy of the ground state for a given geometry [36, 37]. This can be intuitively understood by recognizing that atomic positions and atomic numbers are encoded in the spatial distribution of the density. Consequently, one can reconstruct a Slater determinant and determine an energy using SCF. This implies that all the information we need is encoded within this 3D function:

$$n(\mathbf{r}) = N_e \int |\Psi(\mathbf{r}, \mathbf{r_2}, \mathbf{r_3}, \dots \mathbf{r_{N_e}})|^2 d\mathbf{r_2} d\mathbf{r_3} \dots d\mathbf{r_{N_e}} \tag{1.3.18}$$

Using a 3-dimensional function as a variable, compared to the $3N_e$ dimensions of the wavefunction, makes this approach more computationally efficient. However, the assumption that $n(\mathbf{r})$ could be utilized in this manner had not yet been demonstrated. This changed when Hohenberg and Kohn introduced their two foundational theorems to the scientific community in 1964 [38]. This event is considered the true birth of Density Functional Theory (DFT). We do not present the detailed proofs here but instead discuss their implications:

Theorem 1: For any system of interacting electrons, there is a bijection between the Hamiltonian and the electronic ground state density $n(r)$, making the total energy of the ground state a functional of the electronic density.

Now we can express the total ground-state electronic energy as a functional of the density:

$$E_e[n(\mathbf{r})] = T_e[n(\mathbf{r})] + V_{e,e}[n(\mathbf{r})] + V_{ext}[n(\mathbf{r})] \tag{1.3.19}$$

With V_{ext} the external interaction with the nuclei. The other two terms T_e and $V_{e,e}$ are the kinetic and Coulombic terms. Their exact expressions are not known, so we will need to make a new guess.

$$V_{ext}[n] = -e^2 \int \sum_{j=1}^{N} \frac{Z_j}{|\mathbf{r} - \mathbf{R_j}|} n(\mathbf{r}) d\mathbf{r} \tag{1.3.20}$$

Theorem 2: The energy of the ground state is given by the global minimum of the energy functional $E[n]$.

According to this second theorem, the minimization of $E[n]$ is sufficient to determine the density of the ground state and the total energy, in accordance with the variational principle.

In a first approximation, the $V_{e,e}$ can be approximated with a continuous "mean field" expression:

$$V_{e,e}[n] = V_{c(e,e)}[n] + C[n] \tag{1.3.21}$$

$$V_{c(e,e)}[n] = \frac{1}{2} \iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \tag{1.3.22}$$

Here, $C[n]$ is a correction term due to self-interaction. There is no direct formula for the kinetic term $T_e[n(\mathbf{r})]$ as it is difficult to determine the velocity of individual electrons at a static density.

The Kohn-Sham equations

In 1965, to approach T_e, Kohn and Sham[39] ref proposed to create a correspondence between the electronic density $n(r)$ and a non-interacting electron system that generates the same density. We indicate with S this fictitious system:

$$E[n] = E_s[n] + E_{XC}[n] \tag{1.3.23}$$

Kohn and Sham proceed by linking the density of the ground state $n_0(\mathbf{r})$ to a set of decoupled spinorbitals: $\{\psi_i(\mathbf{r})\}$. These orbitals correspond to the density by a unique functional of the density thanks to the Hohenberg and Kohn theorems.

$$n(\mathbf{r}) = \sum_{i=1}^{N_e} \langle \psi_i | \psi_i \rangle \tag{1.3.24}$$

By doing this, we reintroduce orbitals into DFT. We can easily determine the corresponding values of $E_s[n]$, using the Hartree-Fock way :

$$E_s[n] = T_s[n] + V_s e, e[n] - e^2 \int \sum_{j=1}^{N} \frac{Z_j}{|\mathbf{r} - \mathbf{R_j}|} n(\mathbf{r}) d\mathbf{r} \tag{1.3.25}$$

With:

$$T_s[n] = -\frac{1}{2} \sum_{i=1}^{N_e} \langle \psi_i[n] | \nabla^2 | \psi_i[n] \rangle \tag{1.3.26}$$

$$V_s e, e[n] = V_c e, e[n] = \frac{1}{2} \iint \frac{n(r)n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \tag{1.3.27}$$

Since $E_{XC}[n]$ is an undetermined corrective term up to this point, Equation 1.3.23 is exact. By definition, E_{XC} contains the exchange and correlation term needed to transition

from the non-interacting to the interacting many-body problem. It also includes the self-interaction correction of the mean field. We must make an approximation to calculate this term, which will inevitably introduce some errors.

Using the Kohn-Sham scheme, we can again split the many-body problem into single-electron Schrödinger equations. These are the Kohn-Sham equations :

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}} + e^2\int\frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' - e^2\sum_{j=1}^{N}\frac{Z_j}{|\mathbf{r}-\mathbf{R_j}|} + V_{XC}(\mathbf{r})\right]\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{1.3.28}$$

Where V_{XC} is the undetermined exchange-correlation potential, formally given by:

$$V_{XC}(\mathbf{r}) = \frac{\delta E_{XC}[n(\mathbf{r})]}{\delta n(\mathbf{r})} \tag{1.3.29}$$

Once again, this set of equations can be solved using a SCF algorithm, with a new intermediate step in which the parameters of the Kohn-Sham equation are determined from the density. A significant advantage is that we obtain molecular orbitals with associated energy levels within a framework where they were not originally calculated. However, the energy levels of these orbitals strongly depend on the exchange-correlation functionals and are based on the ground state density.

1.3.4 Exchange and correlation functionals

Once we have expressed in detail the main aspects of DFT, we have to present how we deal with the exchange-correlation functionals. The chosen approximation has a significant impact on the reliability of the results. They are usually classified by complexity on a Jacob's ladder[40, 19] (see Figure 1.3.1).

In this thesis, we will briefly discuss three types of E_{XC} from the surface: Local Density Approximation (LDA), the Perdew-Burke-Ernzerhof (PBE) functional, and the hybrid functional PBE0.

LDA: The straightforward approximation assumes that the exchange-correlation functional can be represented as an integral over the entire space of the exchange-correlation energy per electron in a homogeneous electron gas, ϵ_{xc}^{hom}:

$$E_{LDA}[n(\mathbf{r})] = \int n(r)\epsilon_{XC}^{hom}(n(\mathbf{r}))d\mathbf{r} \tag{1.3.30}$$

This ϵ_{xc}^{hom} is determined using accurate quantum Monte Carlo simulations [41, 42, 43]. This approximation, one of the earliest introduced, is effective in calculating the properties of solids with electronic densities similar to a homogeneous electron gas, particularly in systems where the density varies slowly in space, such as crystals, especially metallic ones [44, 45] . However, it struggles to accurately reproduce properties of isolated molecules or systems with strongly correlated electrons. It also overestimates hydrogen bonding [46]

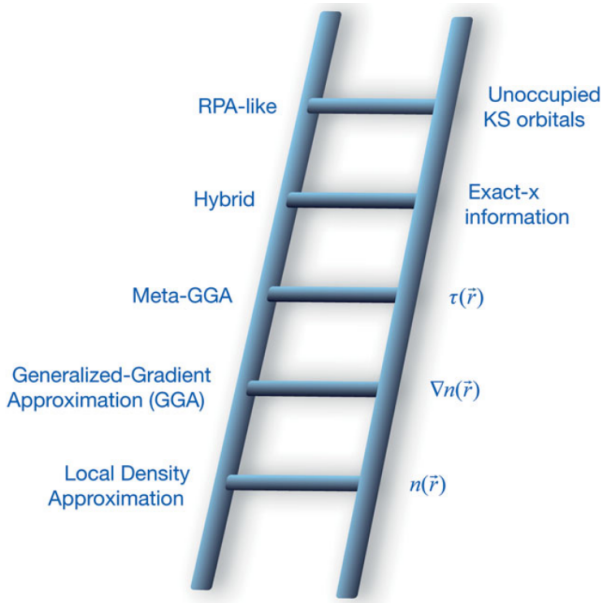


Figure 1.3.1: The Jacob's ladder proposed by Perdew in 2013[40], with name of the family of E_{XC} on the left and added tools to calculated them on the right

. LDA is not a good choice to study chemical reactions in water, where electrons are localized in molecular orbitals and where hydrogen bonding plays a crucial role.

PBE: The second-level approximation also includes the local gradient of the density. This functional family is known as the Generalized Gradient Approximation (GGA) family.

$$E_{XC}^{GGA}[n(\mathbf{r})] = \int \mathrm{d}r\ n(\mathbf{r})\, \epsilon_{XC}^{GGA}(n(\mathbf{r}), \nabla_{\mathbf{r}} n(\mathbf{r})) \tag{1.3.31}$$

This is the case of the PBE functional [47]. For PBE, to accurately take in account gradient effects, $\epsilon_{XC}^{GGA}(n(r), \nabla_{\mathbf{r}} n(\mathbf{r}))$ is fitted on the behavior of a electron gas with a constant gradient: $n(\mathbf{r}) = n(\mathbf{r_0}) + (\nabla_{\mathbf{r}} n(\mathbf{r_0}))(\mathbf{r} - \mathbf{r_0})$.

As PBE has no external parameters other than those defining ϵ_{XC}^{GGA}, which are not fitted to other observations, it is *ab initio*. This functional is considered useful because most higher levels of approximation are considered at least semiempirical by some researchers in the domain[40]. However, in the case of chemical reactions, PBE has a significant drawback: it underestimates the energy of geometries with delocalized electrons (sometimes by more than 10 kcal/mol [48]). Often in chemical reactions, these delocalized geometries correspond to transition states in which electrons are moving to break and/or form covalent bonds. This means that using PBE will be less accurate for energy barriers than for reaction gaps, making the treatment of kinetics less reliable. However, since it is currently one of the few affordable *ab initio* functionals for molecular dynamics, we will use it while keeping its intrinsic limitations in mind.

PBE0: hybrid functionals include a portion of the Hartree-Fock exchange (see Eq 1.3.17). This makes them non-local but also reduces some of the computational advantages of DFT. The general idea is to average between the GGA exchange functional (E_X^{GGA}, obtained by separating the exchange and correlation parts of the XC functional) and the exact Hartree-Fock exchange calculated using the Kohn-Sham orbitals.

Among all commonly used hybrid DFT methods, PBE0 is the one that can most plausibly be considered *ab initio* [49], even if this is not the opinion of all researchers. PBE0, which is based on PBE, does not introduce any external parameters beyond those of the GGA functional. Furthermore, PBE0 hybridization is achieved through a fixed coefficient that relies on a marginal approximation of the adiabatic connection formula[50].

$$E_{XC}^{PBE0} = E_{XC}^{PBE} + 1/4(K_X^{HF} - E_X^{PBE}) \tag{1.3.32}$$

Despite this simple hybridization, PBE0-DFT demonstrates its reliability to the rest of the community and "seems to be a good compromise for those who want to obtain fair accuracy for systems ranging from molecules to solids and at the same time have a direct connection to physical principles", according to Ernzerhof and Scuseria [51].

However, the software we used for this thesis, CPMD[52], took 20 times longer to perform PBE0 calculations compared to PBE, even with an educated guess for the initialization of the wavefunctions (see Appendix C). The use of this hybrid functional is a step towards accuracy for the *ab initio* studies of chemical reactions in solution. This issue would likely be addressed by increasing computational power or using a hybrid trained Machine Learning Inter-atomic Potential (MLIP). We will present MLIP in more detail in the following subsection.

The accuracy of how we calculate the forces is not the only challenge in this domain. In the next chapter, we will define the specificities of studying chemical reactions in condensed phases and the issues associated with these types of transformations. However, our discussion on forces estimation is not yet complete; some technical features that have their own importance need to be added. We also need to briefly introduce MLIPs.

1.3.5 Computational aspects

Atomic and Plane Wave basis sets

When performing calculations on a computer, especially when there are no analytical solutions, as is the case for quantum mechanics, the use of a discrete decomposition for functions we want to optimize is mandatory. To perform the optimization of the wavefunctions correctly, we decompose them into a finite basis set of 3-dimensional scalar functions and then optimize the parameters of the linear combination they form.

The most widely used basis functions are Gaussian functions, centered on atomic nuclei and multiplied by polynomials [19]. One of the significant advantages of Gaussian

functions is that the Coulomb terms of Hartree–Fock calculations can be computed analytically, resulting in much faster computations. This advantage also extends to hybrid-DFT calculations. However, there are two major drawbacks to using Gaussian basis sets. Firstly, they are not naturally suited for periodic box simulations (see subsection 1.1.2). Secondly, there is a finite set of possible Gaussian basis sets (even though libraries are regularly updated). Regardless of the software used, one will eventually reach the maximum possible decomposition.

Another approach, with its own advantages, is to use plane-wave (PW) basis sets to decompose the wavefunction into its Fourier components, taking advantage of its periodicity. Each plane wave in the base is associated with a kinetic energy: $E = |G|^2\hbar^2/2m_e$, where G is the wave vector of Plane Wave (PW). A cutoff point can be set for this kinetic energy, beyond which the expansion is terminated. A higher cutoff result in a more accurate decomposition. Thus, while plane waves are less effective for Hartree-Fock exchange, where Gaussian basis sets excel, they are well suited for PBC and can achieve the desired precision in the decomposition by increasing the kinetic energy cutoff to an appropriate value.

From a theoretical point of view, PW basis sets are as accurate as we need, are easier to manipulate and present good properties with periodic systems like solids. In the case of liquids the periodic boundary conditions make them primitively periodic even if it is not the case for real liquids. For these reasons, they are generally preferred by physicists. However, Gaussian basis sets are often more efficient, particularly for Hartree-Fock exchange, which is crucial for accurately describing molecules. Typically, chemists prefer Gaussian basis sets, especially for calculations in the gas phase where PBC are not required.

More recently, codes such as CP2K [53] proposed a Gaussian Plane Wave (GPW) basis set: a dual basis of Gaussian orbitals centered on atoms and PW[54].

In this manuscript we only use PWs, as the systems we study are liquids with PBC conditions and because up to now we do not have to perform hybrid-AIMD, so we do not need the Gaussian efficiency. However, part of the technical work performed during this thesis was to pave the way for changing the environment we used from PW basis, using the CPMD software[52], to the GPW of CP2K, in future works. This transition offers us a more regularly updated software along with the possibility to calculate Hartree-Fock exchange.

Pseudo-Potentials

When studying chemical reactions, valence electrons typically have the most significant impact on the behavior of the system. Core electrons, on the other hand, contribute to the computational cost of the model without significantly modifying the reaction process. This cost is all the more important because the core electron wave functions have strong and short oscillations near the nuclei that make them very difficult to decompose on the

PWs, demanding unnecessarily high cut-offs.

The idea behind pseudopotentials is to replace all the core electrons and the effect of the nuclei by a screened effective potential capable of masking the fast oscillations near the atomic centers and reproducing the true external potential once a cutoff ray is reached. It is crucial to select the appropriate pseudopotentials for the type of atom and the specific DFT method that we use. In this thesis, we used the Martins-Troullier pseudopotentials parameterized by PBE [55].

Van der Waals (VDW) empirical corrections

If the use of pseudopotentials can be considered as *ab initio* because it is based only on the full electronic results of the same DFT method, this is not the case for the parameterized potential introduced here.

Van der Waals (VDW) interactions come from the polarization and / or polarizability of the electronic clouds. They combine into an attractive potential proportional to $-d^{-6}$ where d is the interatomic distance. DFT methods struggle to efficiently represent this interaction because they are local or almost local methods and VDW forces are based on long-range electronic cohesion effects.

To efficiently represent the true behavior of our systems, especially for the physical properties of the solvent, an empirical correction of the potential adding the $-d^{-6}$ term is usually needed. This adjustment compromises the true *abinitio* forces, but alternatives are limited. We can expect that these corrections do not have a strong impact on barrier heights and reaction energy gaps because they mainly rely on covalent bounding. However, it is hard to estimate the influence that it has on the kinetics.

In this work, we used the D2 version of the Grimme's VDW empirical corrections [56].

1.4 Machine Learning Interatomic Potentials (MLIP)

A significant part of the results presented in this work employ machine learning potential methods. This highlights how these methods have become a cornerstone in the current landscape of *ab initio* calculations.

MLIPs encode within their internal parameters the observed behavior of another potential (here DFT). They can overcome two limitations of DFT:

- The computational cost of MLIPs is one to three orders of magnitude lower than that of DFT. This allows for more accurate sampling using longer calculations or for calculations of comparable size using reduced computational resources.

- The scaling factor of MLIPs is linear with the size of the system. This allows us to expect Machine Learned Molecular Dynamics (MLMD) to become increasingly efficient relative to AIMD as the complexity of the system increases, particularly if MLMD potentials can be applied to larger systems than those used to train them.

In this work, MLIPs are generated using a neural network and are referred to as Neural Network Potentials (NNPs). A neural network is a computational model consisting of interconnected "neurons" organized in layers. It processes input data, adjusts internal weights for each neuron's connections during training, and is used for tasks such as pattern recognition and classification. In our context, it is used to calculate the energy of a given geometry.

We can sample the phase space using a combination of an AIMD dataset, generated on demand, and home trained MLIPs [57]. This approach is referred to as MLMD, in contrast to full AIMD. Since MLMD is an aspect of this thesis that does not involve development, the following description focuses on its application.

1.4.1 Training

In 2007, Behler and Parrinello [58] proposed to generate a NNP, using an atomistic decomposition of the energy.

$$E_{tot} = \sum_{i=1}^{N} \epsilon_i \tag{1.4.1}$$

where the $\{\epsilon_i\}$ are calculated using the neural network applied on a determined environment centered on the atoms. Rotation, translation, and permutation invariances of the potential, necessary to respect the laws of physics, are guaranteed by the use of appropriate descriptors. Technically, descriptors are CVs, but of a high level of complexity. They are trained or designed to capture the local atomic environment from Cartesian coordinates and to represent it as vectors that are invariant to rotation and permutation. These descriptors act as filters between the Cartesian coordinates and the neural network.

To train a NNP, we compare its predicted results for a given geometry with the correct answer from the dataset. Technically, the only output we obtain from the NNP is the total energy, since the NNP atomic energy decomposition is not necessarily physically accurate. However, we can still obtain the forces by estimating the gradient of the output energy with respect to the coordinates. This is feasible because every function within the network is differentiable.

Once the NNP energies and forces have been calculated for a set of parameters $\mathbf{w}$, we can define a loss function:

$$\mathcal{L}(\mathbf{R}) = \frac{1}{N_{geom}} \sum_{i=1}^{N_{geom}} \left[(E^i_{NN} - E_i)^2 + \frac{\beta}{3N} \sum_{j=1}^{3N} (F^i_{jNN} - F^i_j)^2 \right] \tag{1.4.2}$$

where N_{geom} is the number of geometries in the training set, E^i_{NN} is the energy computed from the neural network, E_i is the true energy in the dataset, F^i_{jNN} is the j-th atomic coordinate of the forces vector from the neural network, and F^i_j is the reference force. β is the balance parameter defining the relative importance of energy and forces. Using

the same property of differentiation previously adopted, we can obtain the gradient of the loss function with respect to its parameters and then perform a gradient descent to minimize the loss function. This improves the accuracy of the predictions across the dataset geometries.

In the following, we will use a loss function similar to the one presented earlier but weighted to account for the heterogeneity and solvent predominance in the simulation.

$$\mathcal{L}(\mathbf{R}) = \frac{1}{N_{geom}} \sum_{i=1}^{N_{geom}} \left[p_E (E^i_{NN} - E_i)^2 + p_f \frac{1}{N_{elem}} \sum_{j=1}^{N_{elem}} \frac{N}{n_i} (F^i_{jNN} - F^i_j)^2 \right] \tag{1.4.3}$$

where, n_i is the number of atoms of type i, N_{elem} is the number of element types and N is the total number of atoms. The weights in the force term ensure that each elements of the system are equally represented. This loss function has proven its reliability in previously published work [57].

1.4.2 The NNP committee

To use MLIPs effectively, caution is required. If we encounter geometries for which the training data lack information or guesses, the predicted energies may become non-physical or even diverge. During dynamics, we will explore geometries and configurations outside of the training dataset, and we may lack information about the accuracy of our NNP for these new geometries. To address this issue, we can train multiple NNPs to form a committee and use this group to assess the quality of predictions by evaluating the uncertainty among the committee members, following the approach of Schran *et al.* in 2020 [59]. This involves training NNPs on the same dataset but with different initial parameters determined by various random seeds. The final energy and forces can be obtained as the mean value between the committees. The standard deviation of the predictions from different committee members can be used as an indicator of prediction reliability. However, there is no way to ensure the absence of systematic errors in the extrapolation. To avoid regions of the phase space where the MLIP may be unreliable and to allow longer sampling, we can use a "mirror reflection" approach, as presented in the following reference[57].

1.5 Conclusion of this chapter

Throughout this chapter, we have thoroughly introduced the theoretical background necessary to perform AIMD, including the dynamics algorithms, the methods to estimate forces, and DFT. We have also discussed key concepts from statistical mechanics that enable us to link the results observed in our simulations to macroscopic phenomena. In addition, we have introduced the concept of MLMD to address some limitations of AIMD.

The tools we have covered are theoretically sufficient to sample the complete phase space of a system and estimate the thermodynamic and kinetic properties of a reaction. However, in the practical case of chemical reactions in solution, this approach alone is inadequate. The energy barriers associated with covalent bonding are extremely high. For a barrier of $30k_BT$, the time required to observe such events spontaneously would be approximately one month in real time. In simulations, with a time step of 0.5 fs and a second of calculation per step, this would equate to approximately 10^{13} years of continuous simulation.

In the next chapter, we will focus on the challenges of studying chemical reactions in solution and present the tools available to overcome these barriers.

Chapter 2

Chemical reactions: coordinates and enhanced sampling

Introduction to time and dimensionality curses

The real-time simulation achievable in AIMD simulations is of the order of 1 ns, which corresponds to approximately 2 million data points. Although MLIPs alleviate this limitation, they are not yet sufficient to overcome reaction barriers, which are often referred to as "rare events". This issue is commonly described as the curse of time.

To sample these transitions and calculate the associated thermodynamics and kinetics, a series of protocols have been developed. These protocols generally fall into two main categories of approaches:

1. **Biased Hamiltonian method**: This approach generally involves changing the energetic background by adding a known bias that depends on CV. What was considered "rare" for an unbiased Hamiltonian could be very common for a modified one. The primary challenge of this method is to recover reliable information about the true Hamiltonian.

2. **Biased initialization method**: This approach is harder to grasp. It involves modifying the sampling by orienting the selection of initialization points for new trajectories using external criteria. In this case, the "bias" is more difficult to quantify as it is a purely statistical bias, in a large sense, due to the selection process.

These methods are explicitly used to modify the sampling and diverge from the equilibrium distribution (see subsection 1.2.2). They are referred to as "enhanced sampling techniques". Typically, when researchers describe a trajectory as "biased", they refer to the first type of methods. In the following sections, we will adopt this restricted definition of "biased", while noting that it is an incomplete description. A more comprehensive discussion will be provided in the results part of this manuscript, particularly when we will be addressing the application of stochastic dynamics equations.

Whether for the addition of external forces or the selection process, we rely at least partially on CVs. The quality of the chosen CVs and their involvement in the reaction mechanism can drastically affect the quality of the obtained data. The challenge arises from the vast number of possible CVs. Identifying which CVs accurately describes the reaction (i.e., is an effective Reaction Coordinate (RC)) is difficult. The main problem is that determining which coordinate is preferable requires extensive data on the reaction process, which are limited because of the "curse of time". Consequently, we must introduce a new ansatz for CV to induce the reaction, study the results, and refine our initial CV. The definition of our ansatz is crucial for the quality of the obtained results. This substantial and unleaded reduction of dimensions - from the $3N$ dimensions of phase space positions to the few CVs used to apply forces to the system- is referred to as the curse of dimensionality.

The present chapter describes in detail how we can reconstruct the microscopic mechanism of a chemical reaction and estimate its thermodynamic properties while overcoming these two curses.

2.1 Construction of variables and reaction coordinates

The search for a final process to generate an accurate RC remains ongoing. When we study chemical reactions, we usually separate our problem into elementary single-step reactions. To work on these elementary processes, we usually define different kinds of CV (see subsection 1.2.3 for definition). To present them, we are going to designate with A the reactants of a given elementary step and with B the products.

From this point on, we shall use the following definitions:

- A CV is a differentiable function of the configuration space.
- An Order Parameter (OP) is a CV that efficiently separates the reactants and the products of the elementary step, which means that the definition regions of A and B are clearly separated and identified using this CV. This variable does not contain systematically information about the mechanism.
- A Reaction Coordinate (RC) is an OP that follows the microscopic mechanism of the reaction. It helps to distinguish each intermediate geometry of the reaction IN THE RIGHT ORDER. It is usually a one-dimensional CV. Ideally, it encompasses all Minimum Free Energy Path (MFEP) and has good dynamic properties (more on this last point in the next chapter).

In the final workflow presented in this manuscript, the steps for defining CVs and performing enhanced sampling techniques are interconnected. We use CVs to perform enhanced sampling simulations and then use the enhanced sampling data to refine CVs. In the following subsections, we provide a detailed step-by-step process for generating

a reaction coordinate. The subsequent section will elaborate on the enhanced sampling methods employed.

2.1.1 Chemical Space of CV

According to the positional component of the equilibrium distribution (see Eq 1.2.7), in the (N,V,T) ensemble, every position in Γ, the configuration space, has a non-zero equilibrium probability. This implies that to explore the configuration space via molecular dynamics, each position within this space should be sampled at least once. In practice, exploring all $3N$ dimensions is impractical. Moreover, many of these degrees of freedom are irrelevant to chemical studies; for example, rotation of a H_2O molecule far from the reactive center is unlikely to significantly impact reactivity. Therefore, we need to focus on a subset of degrees of freedom deemed relevant for chemical reactions in solution. In this manuscript, these selected CVs are referred to as the chemical space, here denoted by $\mho$. The dimension of the chemical space is denoted by N_{chem}, and an element inside it is denoted by $\mathbf{X}$. The definition of $\mho$ is critical for the subsequent parts of our study.

We will present two example systems: a synthetic two-dimensional potential and a realistic chemical system.

The Müller-Brown potential:

The Müller-Brown potential [60] is a synthetic 2D potential V defined as:

$$V(x,y) = \sum_{i=1}^{4} D_i exp[a_i(x-X_i)^2 + b_i(x-X_i)(y-Y_i) + c_i(y-Y_i)^2] \tag{2.1.1}$$

where $D_{1...4} = [-200, -100, -170, 15]$, $a_{1...4} = [-1, -1, -6.5, 0.7]$, $b_{1...4} = [0, 0, 11, 0.6]$, $c_{1...4} = [-10, -10, -6.5, 0.7]$, $X_{1...4} = [1, 0, -0.5, -1]$, and $Y_{1...4} = [0, 0.5, 1.5, 1]$.

This potential has been used in the literature as a benchmark for evaluating the performance of dynamic simulation tools [61, 60, 62].

As illustrated in Figure 2.1.1, this system consists of two nonharmonic wells separated by a twisted barrier. This simplified potential serves as an approximation of the complexity that a true effective potential might exhibit in a multidimensional system. Here, there are only two degrees of freedom, so the chemical space can be assimilated into the full configuration space, $\mho = \{x, y\}$. To sample this potential, we will employ an under-damped Langevin integrator, which is described in the following chapter.

The 5-Methylhydantoin

5-Methylhydantoin (hereafter referred to as hydantoin) is a molecule of significant interest in the study of the origin of life [11].

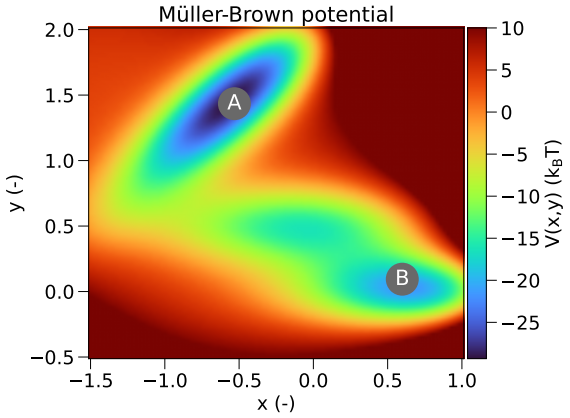


Figure 2.1.1: The Müller-Brown synthetic potential surface with its two main wells: "A" for the global minimum of the surface and "B" for the second. $k_{\mathrm{B}}T = 5$ kcal.mol^{-1}.

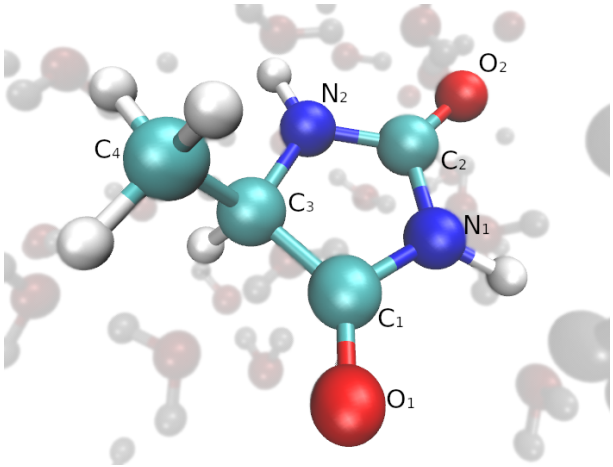


Figure 2.1.2: The hydantoin system, with one hydantoin molecule, one NaCl ion pair, one NaOH ion pair and 97 water molecules

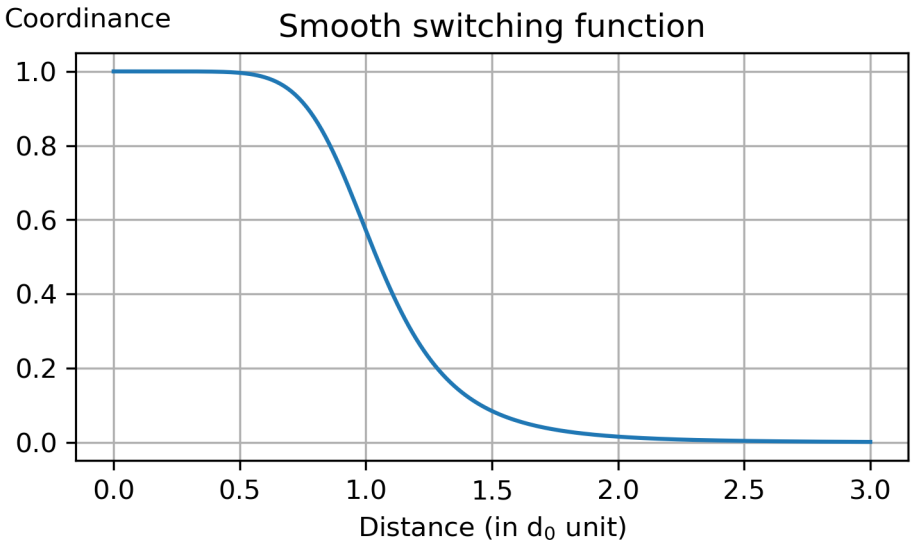


Figure 2.1.3: The smooth switching function in reduced units, with $n = 8$ and $m = 14$

To study its chemical behavior, we can monitor the environment of selected atoms, particularly by focusing on the number of covalent bonds that these atoms form. It is important to understand that the evaluation of the chemical space of a given geometry should be significantly less computationally expensive than DFT by several orders of magnitude. Otherwise, it would be too expensive to perform enhanced sampling. To keep calculations at a relatively low cost, covalent bonding is reduced to distance monitoring, allowing the calculation of the coordination number of a specific atom for a specific element. This simplification of electronic behavior is not problematic because the physical behavior of the system is ensured by the use of DFT to calculate energies and forces. However, it has important consequences that we will include in our methodology.

The formation of a chemical bond is determined using a smooth switching function (see figure 2.1.3). The sum of an atom i and a family of elements σ is the coordinate number of the atom i in the σ element: $\mathrm{c}_i(\sigma)$.

$$\mathrm{c}_i(\sigma) = \sum_{j \in \sigma} \frac{1 - \left(\frac{d_{ij}}{d_0}\right)^n}{1 - \left(\frac{d_{ij}}{d_0}\right)^m} \tag{2.1.2}$$

where d_0 depends on the two elements involved in the bonding and is typically chosen based on the radial distribution function of the pair of elements. The parameters n and m are selected so that $m > n$, with typical values of $m = 8$ and $n = 14$. The parameter d_0 does not correspond to the typical length of the bond but rather to the cleaving distance or, more precisely, the transition distance where the bond is half-formed. This definition of a bond does not differentiate between single, double, or triple bonds significantly; Due to the shorter lengths of double and triple bonds, their contributions are simply closer to one in the coordination. This can cause issues when monitoring reactions involving these

types of bonds.

Once we have established the different values d_0 for our system, we can follow the coordination numbers of a set of atoms of interest, $\{A^i\}$, for every element σ_j in the box. This can be represented as a matrix X of shape (i, j). This matrix defines our chemical state [63]. The following is the case for hydantoin:

$$\mathrm{X} = (\mathrm{c}_{A_i}(\sigma_j)) = \begin{pmatrix} & C & O & N & H \\ C^1 & 0.9 & 1.0 & 1.0 & 0.0 \\ C^2 & 0.2 & 1.0 & 1.7 & 0.0 \\ C^3 & 1.6 & 0.1 & 0.9 & 0.7 \\ C^4 & 0.8 & 0.0 & 0.1 & 2.1 \\ N^1 & 1.8 & 0.2 & 0.1 & 0.6 \\ N^2 & 1.9 & 0.1 & 0.1 & 0.7 \end{pmatrix} \tag{2.1.3}$$

The number of atoms corresponds to the representation in Figure 2.1.2. It is crucial to emphasize the selection of nuclei that we follow in this context. As stated in chapter 1, this selection implies that our approach is not completely agnostic. The choice of nuclei is primarily based on general chemical knowledge and a trial-and-error process. The latter case can become computationally expensive.

In practical cases, the number of dimensions, N_{chem}, typically varies between 4 and 30.

2.1.2 Path Collective Variables

The definition of the chemical space reduces the number of dimensions of the problem from $3N$ to N_{chem}. However, most methods used to sample rare events become unaffordable when based on CVs greater than 2 dimensions. To perform this final reduction of dimensionality, we use Path Collective Variables (PCVs).

PCVs are couples of 1 dimensional CVs, denoted s and z. They were introduced by Branduardi *et al* in 2007 [64] as a tool to explore complex free energy in an intuitive way:

$$\begin{cases} s(\mathrm{X}) & = \dfrac{\sum_{\alpha=1}^{N_{ref}} \alpha \times \exp(-\lambda \mathrm{D}[\mathrm{X}, \mathcal{X}_\alpha])}{\sum_{\alpha=1}^{N_{ref}} \exp(-\lambda \mathrm{D}[\mathrm{X}, \mathcal{X}_\alpha])} \\ \\ z(\mathrm{X}) & = \dfrac{-1}{\lambda} \ \ln\left(\sum_{\alpha=1}^{N_{ref}} \exp(-\lambda \mathrm{D}[\mathrm{X}, \mathcal{X}_\alpha])\right) \end{cases} \tag{2.1.4}$$

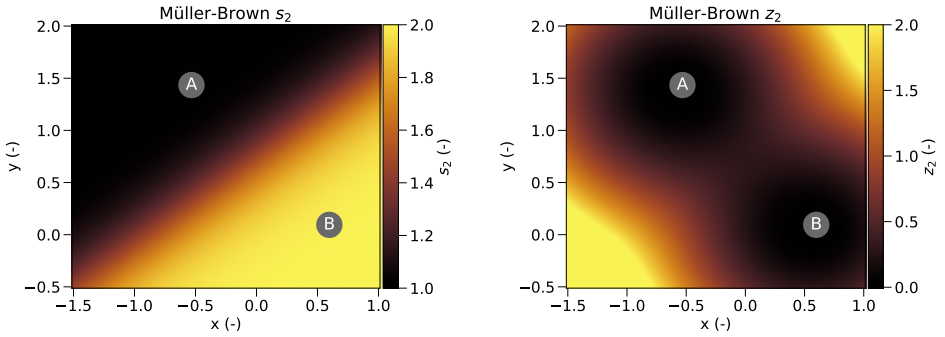
where D is a distance associated with the chemical space, and $\{\mathcal{X}_\alpha\}$ are N_{ref} fixed reference position chosen in the chemical space.

λ is an external parameter defined as:

$$\lambda \times mean(\mathrm{D}[\mathrm{R}^\alpha, \mathrm{R}^{\alpha+1}]) = -ln(0.1) \approx 2.30 \tag{2.1.5}$$

s and z are complementary. They project the chemical space into a 2D space that depends on a designated path:

- s can be seen as a weighted average value of the indices α of the reference structures. The weights, $\{\exp(-\lambda \mathrm{D}[\mathrm{X}, \mathcal{X}_\alpha])\}$, are designed to emphasize the smallest distances.



(a) (b)

Figure 2.1.4: The behavior of s_2 and z_2 on the 2D surface of the Müller-Brown set of variable. "A" denotes the global minimum of the Müller-Brown potential and , "B" is the second one. **(a)** the behavior of s_2 that discriminate the two minima. **(b)** the behavior of z_2 that indicate the two references of the path.

Consequently, s approximates the index of the closest reference structure: if $\mathrm{X} \approx \mathrm{X}_\alpha$, then $s \approx \alpha$. This means that if a trajectory in the configuration space, (R_n), is associated with a trajectory in the chemical space, (X_n), which follows the reference set from $\mathcal{X}_1$ to $\mathcal{X}_{N_{ref}}$, then s progresses from 1 to N_{ref} along the trajectory.

→ If the reference set is chosen to follow the evolution of a chemical reaction, then s indicates the stage of the reaction. Thus, it is a good candidate for a reaction coordinate.

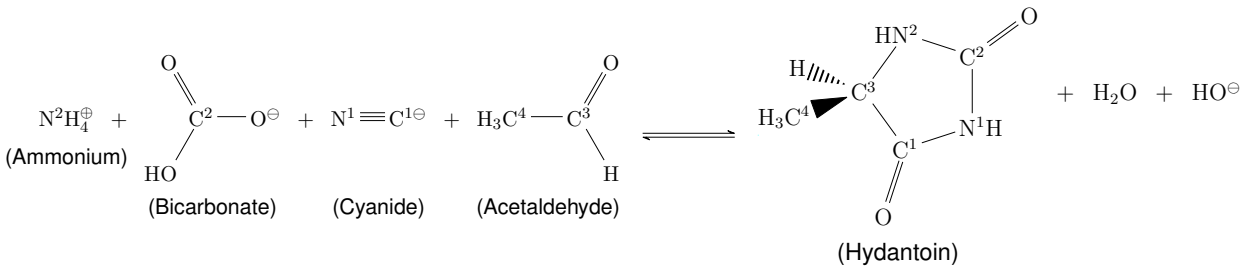
· z represents the sum of all distances between X and the set $\{\mathcal{X}_\alpha\}$. For this reason, it is often referred to as the distance variable.

→ If the z variable is close to 0, the observed geometry is close to the predicted path. In contrast, if z increases, it indicates that we are exploring parts of the chemical space not included in the reference set.

Usually s and z are denoted using the number of references inside the set. For example, s_2 and z_2 designate PCVs defined using two references. To concretely represent what a PCV is, we define a pair (s_2, z_2) for both of our previously introduced examples.

The Müller-Brown potential: two-references path

The two main minima in the Müller-Brown potential are located at $(-0.58, 1.39)$ and $(0.55, 0.05)$ in x and y. To create an initial pair of PCVs, we can use these two positions as the first reference set. The resulting s_2 and z_2 are depicted in Figure 2.1.4. The behavior of the PCVs is consistent with their descriptions: s_2 discriminates between states A and B, and z measures the distance of the position of (x, y) from both A and B.



(a)

$$\mathcal{X}_1 = \begin{array}{c} \\ C^1 \\ C^2 \\ C^3 \\ C^4 \\ N^1 \\ N^2 \end{array} \begin{array}{c} \begin{array}{cccc} C & O & N & H \end{array} \\ \left(\begin{array}{cccc} 0.0 & 0.1 & 1.0 & 0.1 \\ 0.0 & 2.8 & 0.0 & 0.0 \\ 0.8 & 1.0 & 0.0 & 0.7 \\ 0.8 & 0.1 & 0.0 & 2.0 \\ 1.0 & 0.1 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 & 2.5 \end{array} \right) \end{array} \qquad \mathcal{X}_2 = \begin{array}{c} \\ C^1 \\ C^2 \\ C^3 \\ C^4 \\ N^1 \\ N^2 \end{array} \begin{array}{c} \begin{array}{cccc} C & O & N & H \end{array} \\ \left(\begin{array}{cccc} 0.9 & 1.0 & 1.0 & 0.0 \\ 0.2 & 1.0 & 1.7 & 0.0 \\ 1.6 & 0.1 & 0.9 & 0.7 \\ 0.8 & 0.0 & 0.1 & 2.1 \\ 1.8 & 0.2 & 0.1 & 0.6 \\ 1.9 & 0.1 & 0.1 & 0.7 \end{array} \right) \end{array}$$

(b) (c)

Figure 2.1.5: The hydantoin synthesis reaction[65]. **(a)** the equation of the reaction. **(b)** the reference for reactants. **(c)** the reference for products. The two references are defined as the mean value of the chemical space when performing unbiased AIMD in the reactants and products states.

With the exception of the positions we introduce for A and B, no information about the true landscape is encoded within this pair. This implies that they cannot be used as a reaction coordinate but rather as an effective order parameter to distinguish between reactants and products. Using this tool, we can define our reactant and product regions more rigorously: $A = \{s_2 \in [1, 1.2]; z_2 < 0.2\}$ and $B = \{s_2 \in [1.8, 2]; z_2 < 0.2\}$. This definition of A and B is decorrelated from the potential, i.e., from the studied system. This implies that it can be applied to any s_2 and z_2 provided that the two references correctly represent the local minima of the potential in the chemical space.

The 5-methylhydantoin: two-reference path

Suppose that we are studying the synthesis of 5-methylhydantoin from ammonium ion, cyanide ion, carboxylic acid, bicarbonate ion, and acetaldehyde. We can define a pair of PCVs to discriminate between the two states: reactants and products.

The creation of the two references for this chemical synthesis is illustrated in Figure 2.1.5. Given that the chemical space in this case has 24 dimensions, a reliable presentation of the full behavior of s_2 and z_2 cannot be generated. However, by definition, the combination of these two variables can be used as a precise OP.

2.1.3 Reaction coordinates, Transition state and Committor

Once we have obtained an OP for the synthesis we wish to study, we need to take further steps. To understand the chemical mechanisms involved in a reaction, a Reaction Coordinate (RC) is mandatory. This RC encompasses the true behavior of the system during the transition by correctly discriminating every intermediate stage. This definition of a RC is very broad, which is why a wide variety of them can be created for the same transition. Significant work has been done to classify them in terms of "quality" [61, 66, 67].

Among the intermediate stages of the reaction that a RC should help discriminate, the most important is the Transition State (TS). By definition, the TS is the highest state of the chemical transformation in terms of free energy. The TS is challenging to identify perfectly in a $3N$-dimensional landscape. To provide a definition of practical use of it, despite the very large configuration space, we can use the committor.

The committor is defined from an OP and a choice of two regions, A and B, using this OP. Designated as $p_{\mathrm{B}}(\mathbf{R})$, it represents the probability of reaching the product region B before the reactant region A [68]. By definition, if $\mathbf{R}$ is inside B, then $p_{\mathrm{B}}(\mathbf{R}) = 1$; conversely, if R is inside A, then $p_{\mathrm{B}}(\mathbf{R}) = 0$. Based on this definition, we can define the TS region of the reaction from A to B as the ensemble of $\mathbf{R}$ for which $p_{\mathrm{B}}(\mathbf{R}) = 0.5$ or within a threshold around 0.5. This intuitively means that a geometry $\mathbf{R}$ in that situation has the same probability of reaching the reactants or products.

The committor is the "perfect" RC because it quantifies the dynamic behavior of the system for all possible geometries. However, it is also described as "meaningless" [68] in the sense that there is no clear link between the committor values and the geometries of the chemical system.

An important drawback of using the committor is the computational expense needed to accurately determine it. To estimate the committor value of a single geometry $\mathbf{R}$, the only efficient method is the Committor Analysis (CA). This method involves launching a large number of trajectories from $\mathbf{R}$ with the Maxwell-Boltzmann distribution of velocities and approximating $p_{\mathrm{B}}(\mathbf{R})$ by the observed frequency of falling into the B region. This estimation is significantly more expensive than the molecular dynamics itself. For classical dynamics, this can be manageable, but for AIMD, estimating the committor is affordable for only a very limited number of geometries, typically near the TS where the value stabilizes the fastest. Consequently,CA is never conducted on the fly during AIMD. Moreover, it cannot be used directly in enhanced sampling simulations. Instead, a regression of the committor values from other CVs can be used [61, 62, 66].

Committor Analysis on the Müller-Brown potential

To illustrate this, we can use the Müller-Brown potential. We have performed a Langevin dynamics simulation with a time step of 5×10^{-4} ps, a friction term of 50 ps^{-1}, and a temperature factor of $k_{\mathrm{B}}T = 5$ ps^{-2}. The masses of the variables x and y are considered

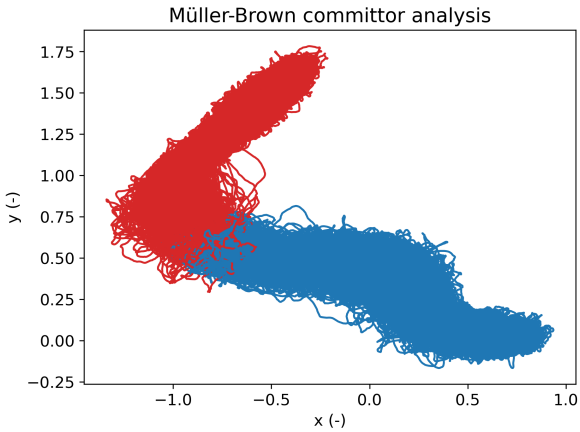


Figure 2.1.6: Committor Analysis (CA) of the (-0.82, 0.62) point of the Muller-Brown potential. The trajectories that commit to B are represented in blue, and in red for A. We can observe an overlapping zone between the two sets of curves near the launching point. This is the zone where recrossing occurs.

equal to 1. We have launched 1000 dynamics simulations with random velocities from the point $(-0.82, 0.62)$ in x and y, which is known as the position of the saddle point between states A and B [60]. The resulting trajectories are presented in Figure 2.1.6.

The final estimation is: $p_{\mathrm{B}}(-0.82, 0.62) = 0.42 \pm 0.03$. This implies that the saddle point identified in the Müller and Brown 1979 paper[60] is not the optimal TS.

2.1.4 Data driven path collective variables

Committor analysis data on a transition state can be used to train new PCVs. The idea is to increase the number of references following a probable transition path. These references must be well-chosen and equally spaced in the chemical space. In the literature, there are some protocols to perform that selection [61, 64, 69]. In this manuscript, we proceed using an algorithm inspired by the Nudged Elastic Band (NEB) method [70]. Let us suppose that we want n references in the final PCV. The two limiting references, R_1 and R_n, are fixed as the average of the chemical space in the reactants and products. The other references are in between.

To clarify the following method, we introduce two notation:

- ➤ The dataset of geometries on which we will search for a path is designated as Θ.
- ➤ The ensemble of all possible paths is designated as $\Theta^{(N-2)}$ because it corresponds to $(N-2)$ intermediate references selected from Θ.

In order to generate an optimized path, we set a fictive potential on $\Theta^{(N-2)}$:

$$E = \sum_{k=1}^{N-1} \left(D_{k,k+1} - \frac{1}{N-1} \sum_{l=1}^{N-1} D_{l,l+1} \right)^2 + \beta \frac{1}{N-2} \sum_{k=1}^{N-2} \left[\max(\theta_{k,k+1,k+2} - \theta_0, 0) \right]^2 \quad (2.1.6)$$

where $D_{k,k+1}$ is the distance in the chemical space between two successive references and $\theta_{k,k+1,k+2}$ is the deviation angle between three successive references.

This potential comes from the work of Théo Magrino[14]. By construction, the global minimum of this potential corresponds to the shortest path with the lowest curvature. To approach this point, we sample this potential using a Monte Carlo process:

1. Select N-2 points at random from the dataset to form the initial reference set: θ^{N-2}. Calculate the associated fictive potential E_0.

2. Pick a new possible reference randomly in the dataset, change with the closest point of θ^{N-2}, and calculate the new potential value E_1.

3. IF $E_1 < E_0$ Validate the new reference set,E_1 becomes the new E_0.

   ELSE Pick a uniform random number a between 0 and 1. If $a < e^{\beta(E_0-E_1)}$, validate the new reference set, E_1 becomes the new E_0.

4. go to 2.

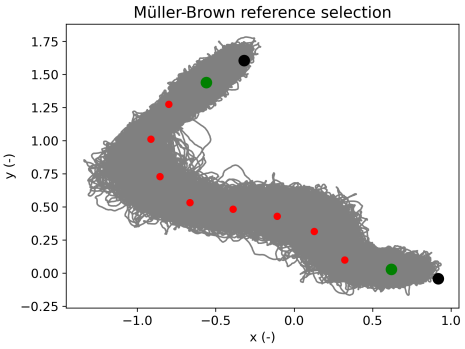
where β is the temperature and is defined as a proportional factor between 0 and 1. The loop is stopped when the variances of the distances and the angles are both below a predetermined threshold.

2.1.4.1 Searching for a highly referenced path collective variables in the Müller-Brown potential

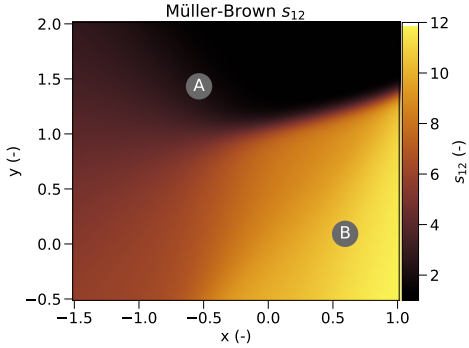
To illustrate the reference search algorithm, we applied it to a dataset composed of the preceding CA trajectories. The results are presented in Figure 2.1.7. Our algorithm successfully generated a path that follows the evolution of the reaction. Now, s_{12} can be used as an estimate RC and z_{12} can be used to monitor the quality of this estimate. Following the same protocol, we can define a RC for any type of chemical reaction, even without prior knowledge of previous work or experiments on the system.

2.2 Enhanced Sampling

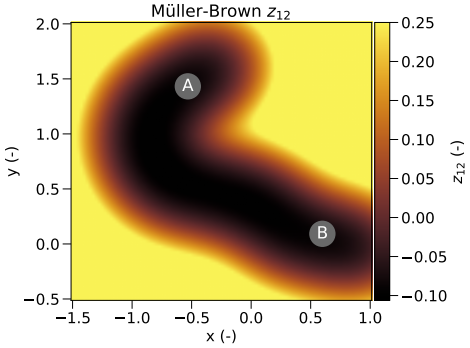
Now that we have defined our working space of collective variables (CV), the chemical space, we can describe the enhanced sampling techniques used to sample rare events. As mentioned in the Introduction, these techniques can be categorized into two types: biased dynamics and oriented initialization. However, instead of introducing these tools based on this classification, we will present them in the natural order in which they are employed in the protocol we developed.



(a)



(b)



(c)

Figure 2.1.7: Results of the research of a PCV with 10 references. **(a)** the validated reference dots in blue and the two fixed dots in green that correspond to the reactant and products. The dataset Θ is represented in grey in the background. The black dots are supplementary references that correspond to linear extrapolation of the path after and before the reactants and the products in order to ensure a correct description of the wells. **(b)** the resulting s_{12} evolving from the reactants to the products and following the shape of the dataset. **(c)** the resulting z_{12} marking the location of the referenced region, we can see it correspond to the dataset location.

2.2.1 Metadynamics

The first tool we employ involves launching a standard molecular dynamics simulation and, at regular time intervals, modifying the Hamiltonian by incrementally adding an external bias. This technique is known as metadynamics [71]. The introduced bias is time dependent and typically takes the form of a sum of Gaussian functions, added every τ_{met} time interval. This bias is applied to a previously defined set of collective variables (CVs). An example using s_2 and z_2 is provided below:

$$V_B(t, s_2, z_2) = a \sum_{k=1}^{t//\tau_{met}} \exp\left\{\left(-\frac{(s_2 - s_2(k\tau_{met}))^2}{2\sigma_{s_2}^2}\right) + \left(-\frac{(z_2 - z_2(k\tau_{met}))^2}{2\sigma_{z_2}^2}\right)\right\} \quad (2.2.1)$$

where $s_2(k\tau_{met})$ and $z_2(k\tau_{met})$ represent the historical values of s_2 and z_2 during the dynamics, evaluated every τ_{met} time interval. The parameter a denotes the height of the introduced bias, while σ_{s_2} and σ_{z_2} are the widths of the Gaussians associated with each variable, typically defined based on the observed distribution in the wells.

A common analogy by Alessandro Laio and Francesco Gervasio [72] illustrates metadynamics effectively. This is the one presented here with small modifications: Imagine a blind person walking randomly on an unknown mountainous terrain. The person can only perceive the immediate area beneath their feet (his current position and the slope of the terrain). As the person is lost and blind, he follows the slope. In a steady state, he wanders around the valley bottom, likely remaining within it. To escape, the blind person can rely on an infinite bag of sand. At regular intervals, he drops sand on his feet, gradually filling the valley. Eventually, the lowest point of the valley reaches the pass level, allowing him to exit. During this process, he has never made any guesses on the location of the pass.

In molecular dynamics, the number of degrees of freedom that we can address by metadynamics is relatively small compared to the total dimensions of the system ($3N$). Typically, metadynamics is applied to up to three collective variables because it is more feasible to "fill with bias" a surface, a volume, or even a 4D hypervolume. The required amount of bias increases exponentially with the number of CVs involved.

Interestingly, if metadynamics is conducted in a sufficiently long simulation time, we can obtain a situation where all possible metastable states of the used CVs are explored and compensated for by the added bias. In this scenario, the system of variables no longer experiences attraction to these local minima and begins to diffuse freely between the metastable states. This implies that the introduced bias approximates the negative of the free energy as a footprint:

$$\lim_{t\to\infty} V_B(t, s_2, z_2) \approx -F(s_2, z_2) + C \tag{2.2.2}$$

where C is a constant term.

However, errors can interfere inside this equation:

- The value of the free energy footprint is always defined up to the height a of the introduced Gaussian functions.
- The accuracy of the free-energy landscape estimation may be compromised if the frequency of bias increments is too high for the system to reach a new steady state. Similarly, if the Gaussian functions are too widely spaced, they may overlap systematically, affecting the results.

These issues highlight that metadynamics represents a trade-off between the quality of the results and the computational cost of the simulation. Reducing both the size of the Gaussian functions and the frequency of increments would probably improve the

estimation of $F(s_2, z_2)$, but these parameters are linearly related to the cost. For example, in Eq 2.2.1, halving a, σ_{s_2}, σ_{z_2}, and the increment frequency simultaneously would result in a sixteen-fold increase in computational cost.

The acceptable size of the parameters depends mainly on the quality of the employed CVs [73], which requires prior information about the system. For this reason, in the protocol described in this manuscript, metadynamics is not used for quantification. Instead, it serves to provide an initial estimate of the transition between reactants and products in a chemical system.

Metadynamics on the Müller-Brown potential

This time, we initiate the dynamics in the vicinity of the metastable state B, as shown in Figure 2.1.1, using the same parameters employed for the committor analysis. In the absence of bias, the system remains stationary near B. To introduce Gaussian functions, we use the collective variables x and y (CVs). Each Gaussian is added every 2000 steps (1 ps) with a height of $1, k_{\mathrm{B}}T$ and a standard deviation of 0.1 for both x and y.

The results of this study are presented in Figure 2.2.1. Although these results, obtained using a synthetic model, exhibit a high quality compared to real molecular dynamics cases, they help us to illustrate the metadynamics process. The results presented in Figure 2.2.1 demonstrate that we achieve staticity (and the zero level of potential) after 270 additions. An initial estimation of this quantity can be made prior to simulation using data from the wells and a guess of the barrier height from empirical knowledge.

In our routine protocol, metadynamics is terminated after the first transition. Typically, it is not feasible to capture the full complexity of the chemical process with only one or two collective variables (CVs) derived from the study of the wells (the only data available without bias). Under these conditions, reaching the diffusion regime would likely require an excessive amount of time, and the final quantification would likely be of little use regarding the quality of the variables.

Having obtained the first transition, we can now refine our results by identifying a transition state (TS) within this transition. Furthermore, we can improve our initial transition using Transition Path Sampling (TPS).

2.2.2 Transition Path Sampling

The primary objective of this process is to refine the transition trajectory we initially obtained by sampling new ones. The goal is to converge towards the MFEP, which represents the path between reactants and products where the underlying gradient of the free energy (with respect to the chemical space) is minimal. Several methods are available to achieve this [13, 74, 75]. An approach involves performing multiple CAs to identify new potential TSs and progressively decorrelate the new trajectories from the initial one. However, a more efficient method is to employ an automatic workflow, such as Aimless

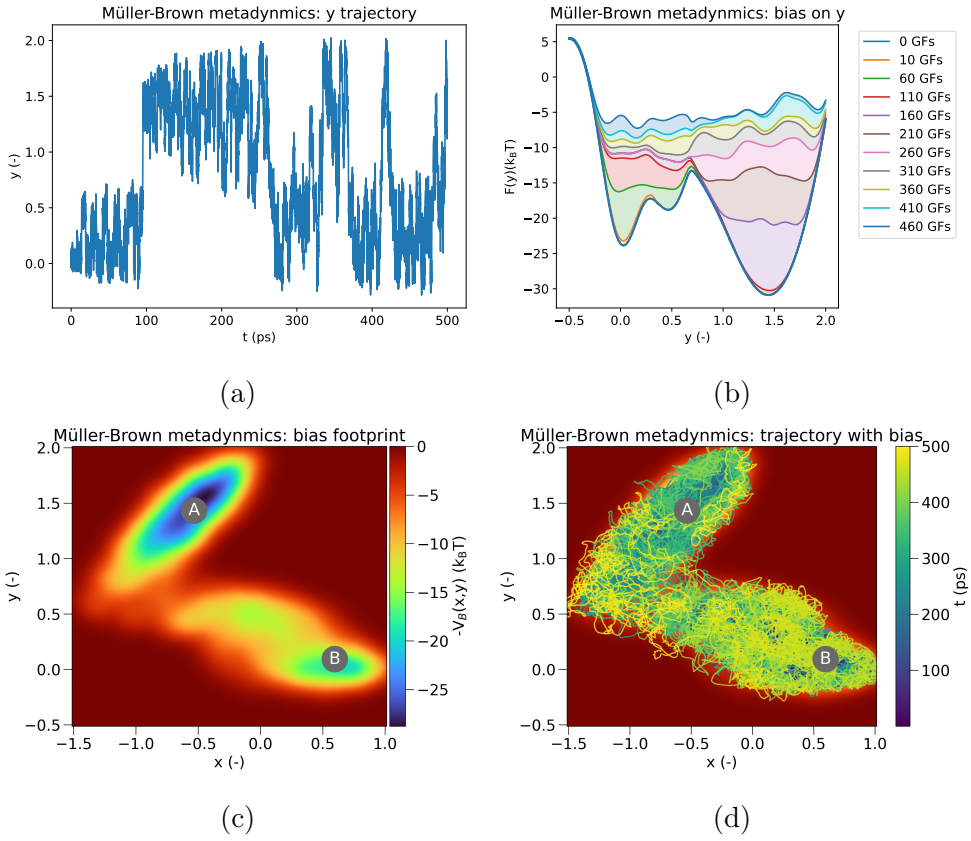


Figure 2.2.1: Results of the metadynamics applied to the Müller-Brown potential. **(a)** The trajectory of y during the metadynamics simulation. We observe the exploration of wells B and A, followed by diffusion after 270 ps. **(b)** The evolution of the bias introduced during metadynamics, projected onto y and superimposed on the actual Müller-Brown potential. This illustrates the "sand" added to the system [72]. The legend indicates the number of Gaussian functions considered for each curve. **(c)** The 2D bias footprint after 500 ps (i.e., containing 500 Gaussian functions). This footprint accurately reproduces the Müller-Brown potential, as shown in Figure 2.1.1. **(d)** The full trajectory overlaid on the bias footprint. Many exchanges between the two potential wells at the tail of the trajectory (highlighted in yellow) suggest that the diffusion regime has been achieved.

Shooting (AS)[76, 77, 78] or Shooting From the Top (SFT) [79].

To initiate a Transition Path Sampling (TPS), similar to the committor analysis, the only required CV is an order parameter Order Parameter to define the regions of reactants and products. Both TPS workflows follow the same algorithm:

1. Take the geometries of an initially obtained transition as a first dataset.
2. Select one candidate transition state geometry from the dataset.
3. Launch 2 trajectories from that candidate with opposed initial speeds drawn from the Maxwell-Boltzmann distribution.

   If One of the trajectories falls into the product region and the other one does the opposite (*i.e.*, we have generated a new transition)

   Then The new transition becomes the new dataset, and the candidate geometry is validated as a potential transition state.
4. Go to 2.

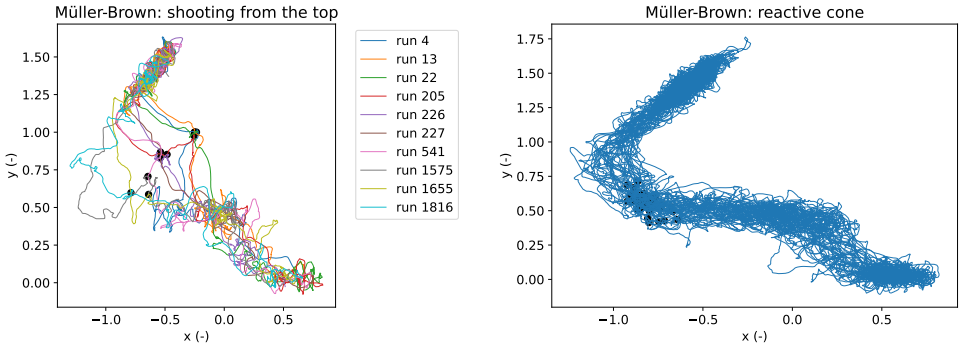
By repeating this process a predetermined number of times, new transitions are sampled, each of which shares a single common point with the previous one. This ensures that each newly validated dataset is progressively more distant from the initial dataset.

The main difference between the two workflows lies in the method used to select the candidate transition states.

For AS, the selected geometry is one of the two geometries located at $\pm\Delta t$ from the last validated transition state on the trajectory of the dataset. This approach has the advantage that the "distance" between the two candidates is a time interval, bypassing geometric considerations. However, since there are only two possible geometries, the workflow may become locked in a situation with low chances of commitment. This issue can be mitigated by selecting a smaller $\pm\Delta t$, although this will slow down the exploration of the transition state ensemble. A balance must be found between the acceptance ratio (the number of accepted TSs relative to the number of attempts) and $\pm\Delta t$ to converge more rapidly towards a new TS.

For SFT, the selected geometry is randomly chosen from all points in the dataset lying within a guessed transition region, which can be defined using an OP. With this method, the distance between two successive guesses is not fixed and can be measured using time intervals or other CVs such as the distance in chemical space. The size and position of the interval can be optimized to achieve the best compromise. This method relies more heavily on the quality of the OP compared to AS.

In both methods, the newly guessed geometries converge toward a saddle point, and the new transition paths converge towards the MFEP associated with this saddle point. If multiple saddle points exist, the system may converge to one of them and remain trapped, in the same way that a metastable state would do in AIMD. At the end of the workflow, if



(a) (b)

Figure 2.2.2: The results of SFT on the Müller-Brown potential. **(a)** The first ten accepted trajectories. We observe the convergence of the accepted points towards the region near the saddle point. **(b)** The second half of the accepted trajectories (the first half is discarded for convergence). This illustrates the formation of the reaction cone around the MFEP. Given that no other degrees of freedom are considered in this simulation beyond those depicted in the figure, and considering that the Müller-Brown potential has only one primary saddle point between A and B, we infer that the SFT has successfully converged.

the initially validated trajectories are discarded for convergence, the remaining validated trajectories will form a "reactive cone" that represents the most probable transition states between reactants and products. This reactive cone effectively represents the intermediate geometries encountered during the reaction mechanism [6]. This forms a dataset typically used to generate highly referenced PCV.

Shooting from the top on the Müller-Brown potential

We performed SFT on the Müller-Brown potential using the same sampling parameters as in the previous two cases. The reactants and products zones are defined as circles centered at the local minima A and B: (-0.58, 1.39) and (0.55, 0.05), respectively, with a radius of 0.2 for both. The transition state region is defined as a circle centered at (-0.82, 0.62), the saddle point identified in the literature [60], with a radius of 0.8. To illustrate the convergence of the process, we initially guessed the transition state at (-0.25, 1.0), which is far from the saddle point location. The results of this study are presented in Figure 2.2.2. The transition path successfully reached the true saddle point region after 10 accepted transitions and remained there once achieved. The acceptance ratio for this study was 0.87%, indicating that the transition region for selection could potentially be reduced because most selected geometries are rejected.

2.2.3 Umbrella Sampling

Once a transition pathway has been obtained with metadynamics and refined with transition path sampling, this information can be used to generate a RC, as described in subsection 2.1.4. The Umbrella Sampling (US), the final step of our protocol, will allow us to accurately evaluate the free energy curve along this RC and consequently the free energy difference between reactants and products [80, 81].

To carry out this procedure, we generate between 15 and 60 copies of our system along the RC, referred to as "windows." Each window is constrained near a specific value of CV with a quadratic and static bias. Each window is sampled using AIMD, and once equilibrium is reached in all windows, the simulation US is concluded. The static bias B for each window is defined as follows:

$$B_i(s) = \frac{k}{2}(s - s_i)^2 \tag{2.2.3}$$

where s is the chosen CV, and s_i is the minimum of the bias introduced in window i. Each s_i is equally spaced, with Δs being the step size between adjacent windows. The strength of the potential k, is determined to ensure adequate overlap, particularly in the case of a flat free energy profile. A common approach is to set the bias to $0.5k_BT$ at the crossing point between two adjacent biases, ensuring that $\frac{k}{2}\left(\frac{\Delta s}{2}\right)^2 = \frac{k_BT}{2}$.

A critical aspect of the US simulation is that the windows should be contiguous in the phase space. This means that transitioning from the data obtained with one window to the others should allow for a progression from the reactants to the products without discontinuities in phase space. If the windows are truly contiguous, the underlying free energy potential for each window, once corrected for the bias, should extend continuously from the potentials obtained with neighboring windows. Consequently, we obtain segments of the free energy curve, one for each window, which can be aggregated together to form the complete curve.

The "aggregation" process of these segments is accomplished using a statistical tool called Weighted Histogram Analysis Method (WHAM). This tool relies on solving the WHAM equations [82, 83, 84]. For a simulation with N_w windows, the WHAM equations are expressed as follows:

$$\begin{aligned} p(s) &= \frac{\sum_{i=1}^{N_w} n_i(s)}{\sum_{i=1}^{N_w} N_i e^{\beta(F_i - B_i(s))}} \\ F_i &= -\frac{1}{\beta} \ln \left(\int_s p(s) e^{-\beta B_i(s)} ds \right) \end{aligned} \tag{2.2.4}$$

where N_i represents the number of points in window i, and $n_i(s)$ denotes the histogram value of s in simulation i at value s. The variables to be optimized are the F_i, the free energy shift applied to each window to re-weight them, and $p(s)$, the final equilibrium

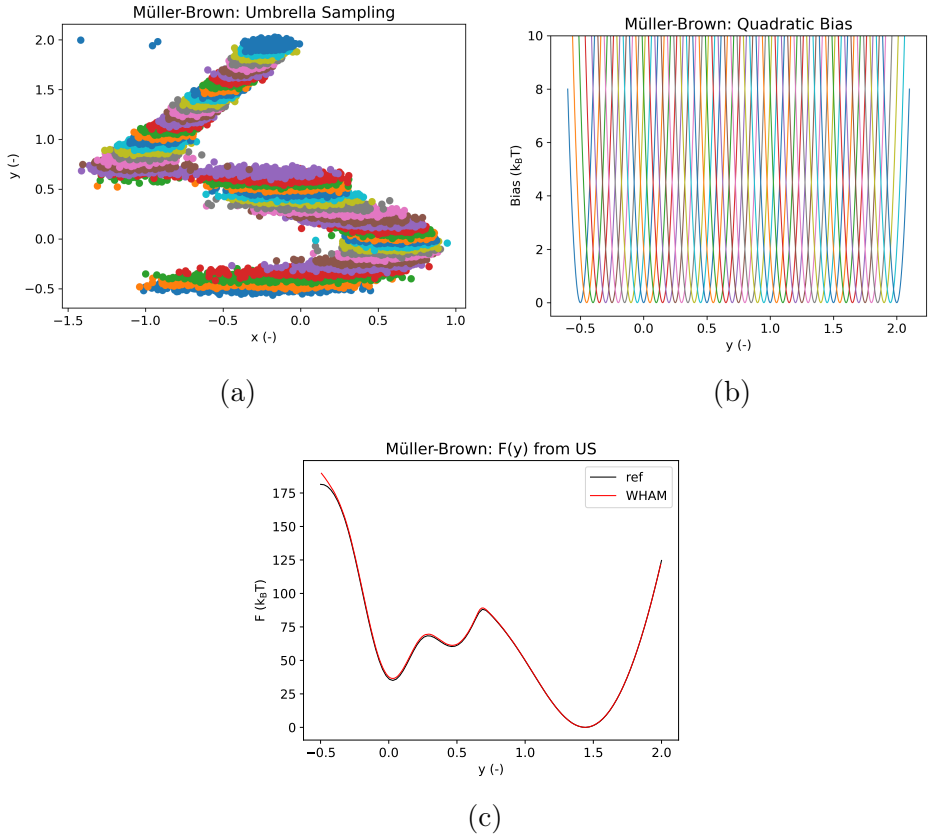


Figure 2.2.3: The results of the US simulation on the Müller-Brown potential. **(a)** The x-y coordinates for each window. Exchanges can be observed at the top of the barrier near $y = 0.6$. **(b)** The bias profiles of the umbrellas along y. **(c)** The reconstructed free energy curve compared with the original potential.

distribution we seek. These variables are determined self-consistently by iteratively solving the two equations.

Umbrella Sampling on the Müller-Brown potential

An umbrella sampling test has been performed on the Müller-Brown potential with respect to the y variable to obtain the underlying free energy curve. This test uses a total of 51 windows, which range from $y_1 = -0.5$ to $y_{51} = 2$, with a step size of 0.05 and a spring constant k of $8000 k_B T$. Each window was sampled during $1 \mu s$. The results of this simulation are presented in Figure 2.2.3. The final free energy curve closely matches the projected potential along y.

Some words about hysteresis

When a gap in the configuration space occurs between two adjacent windows, this indicates that the geometries of these windows are separated by a barrier perpendicular to the reaction coordinate used. This separation can result from an overly long equilibration

of a solvent degree of freedom or from an intermediate step in the chemical process not taken into account in the RC.

In the first case, the slow solvent reorganization and equilibration can be forced by the addition of a static bias. For instance, this situation arises when a spectator amine equilibrates with the water solvent: the characteristic time for protonation is of the same order of magnitude as the available *ab initio* length of the trajectory (≈ 15 ps). Consequently, the reaction may occur randomly in one window but not in others, creating a separation in the configuration space and altering the bulk composition. The time required for this process to reach equilibrium exceeds the available simulation time. Therefore, the preferred solution is to bias the $c_N(H)$ coordination in order to keep this degree of freedom fixed.

The second scenario is termed hysteresis. Here, a portion of the reaction is not captured by the reaction coordinate (*, i.e.*, the reaction coordinate is imperfect). This part of the reaction occurs through degrees of freedom not explicitly considered, and this generally involves a significant barrier. Thus, between these two windows, there exists a hidden barrier and an unmeasured free energy gap, which is not quantified in the WHAM results, as the necessary data are absent from the dataset. Typically, the presence of hysteresis can alter the final results by approximately 10 $\text{kcal} \cdot \text{mol}^{-1}$. This is one of the main sources of error in US simulations.

2.3 Conclusion of this chapter

We have introduced many tools necessary for the thermodynamic study and characterization of chemical reactions in solution. Specifically, we have discussed the concepts of chemical space, committor, and data-based PCV. These elements are fundamental for understanding the behavior and properties of chemical systems at the molecular level.

In addition, we have provided a quite comprehensive overview of enhanced sampling techniques. These methods are essential for obtaining an initial transition state and refining it to achieve a more accurate representation of the system's behavior. This allows us to determine the free energy curve of a collective variable, providing insights into the thermodynamic properties and stability of the chemical system.

Furthermore, we have outlined the datasets available for chemical systems. These data sets include information generated through AIMD, which provides detailed insight into the system's wells, and data obtained from enhanced sampling simulations, which offer a deeper understanding of the system's behavior during the transition.

We can now investigate *terra incognita* of this work by introducing the tools available for the study of the kinetics of dynamic processes in the next chapter. We will explore how this valuable data and these CVs can be processed to investigate the kinetics of chemical reactions.

Chapter 3

Langevin Dynamics: friction and memory for kinetics

3.1 Introduction to coarse-graining

A complementary strategy to address the curses of time and dimensionality is coarse-graining. The idea is to aggregate the behavior of particles into larger sets to work with a simplified representation. There are two ways to define a coarse-grained model.

. Define pseudoatoms that approximate the behavior of groups of atoms, such as water molecules or amino acid residues in a protein. The formed pseudoatoms express the properties of the groups they replace: their mass, their charge, dipole moment, etc.

. Aggregate the behavior of multiple particles into a "mean field" effect on the remaining ones. As the positions of the removed particles are no longer estimated, their influence on the remaining ones becomes random, which reflects our ignorance of their exact positions and velocities.

The first case is used mainly in the field of classical dynamics. In the context of chemical reactions, the variety and complexity of the studied potential make this kind of tool poorly transferable. They are particularly useful in scenarios where there is no covalent bond formation or cleavage, for example, in protein folding.

The second method has historical significance; it is the method that allowed us to understand the Brownian motion of molecules, a milestone in atomic theory. This kind of model allowed us to first estimate the value of the Avogadro constant, thanks to the work of Jean Perrin [85]. This field has gathered the contributions of Einstein [86], Boltzmann [87], Langevin [88], and other physicists of the twentieth century.

Langevin, in particular, developed his approach with the objective of explaining Brownian motion using a stochastic differential equation:

$$m\ddot{x} = -6\pi\mu a\dot{x} + X \tag{3.1.1}$$

where x is the direction of motion, m is the mass of the particle experiencing the motion, and μ is the viscosity of the solvent from the Navier-Stokes equation [18]. X represents the effect of solvent fluctuations, "is indifferently positive and negative, and [...] its magnitude is such that it maintains the agitation of the particle" (translated from French by Lemons and Gythiel in 1997) [88].

This textbook definition shows that the mathematical tools used for the study of random variables have not been developed yet. In more modern terms, this implies that the effects of the solvent can be interpreted as a friction term proportional to the velocity of the studied particle and a random noise with an amplitude linked to the temperature. This is a Markovian process; the movement of x is no longer deterministic but depends on the behavior of a random noise. This kind of stochastic equation of motion is what we call the Langevin equations.

In the following, we will present the different families of Langevin models and describe their properties in detail. To do this, we use a single one-dimensional degree of freedom of the system with a constant mass, stated as x. This choice simplifies the equation and helps to understand the underlying physics.

3.2 Langevin dynamics

3.2.1 Underdamped model

The underdamped Langevin dynamics is the historical approach. It is based on Newton's equation, which means that the position-dependent term of the equation, the mean force, is applied to the acceleration:

$$m\ddot{x} = f(x) - \gamma m\dot{x} + \xi \tag{3.2.1}$$

where m is the mass of x, f is the mean force applied to x, γ is the friction coefficient, and ξ is the noise.

f is directly linked to the free energy curve associated to x:

$$f(x) = -\Delta_x F(x) = k_B T \Delta_x \ln(p_{eq}(x)) \tag{3.2.2}$$

where $F(x)$ is the Helmholtz free energy projected on x (see subsection 1.2.3), and p_{eq} is the equilibrium distribution of x.

The friction coefficient term γ has the dimension of the inverse of time. $1/\gamma$ represents the characteristic response time of the friction term to a perturbation. It is also linked with the auto-correlation of the velocity C_v:

$$C_v(\tau) = \frac{\langle v(0)v(\tau)\rangle}{\langle v^2\rangle} \tag{3.2.3}$$

which has the shape of an oscillating exponential decay with $1/\gamma$ as the characteristic time [89, 90].

The noise has a mean value of 0 by definition: $\langle \xi \rangle = 0$. To respect the fluctuation-dissipation theorem [91, 92], it is an uncorrelated noise with an amplitude of $\sqrt{2\gamma m k_B T}$:

$$\langle \xi(0)\xi(\tau) \rangle = 2\gamma m k_B T \delta(\tau) \tag{3.2.4}$$

This type of decorrelated noise is designated as "white" noise. In practical cases, ξ is usually considered as a Gaussian white noise, as it is the consequence of a large number of particle collisions:

$$\xi = \sqrt{2\gamma m k_B T} W \tag{3.2.5}$$

where W is a Gaussian white noise with an amplitude of 1.

The underdamped Langevin equation, although very simple, is based on two assertions:

- The characteristic time of the fluctuation of the solvent is *shorter* than the timescale resolution of the model.
- The characteristic time of the friction with the solvent is *longer* than the timescale resolution of the model.

This type of model fits well with the case of Brownian motion, where the particles are still under the effect of the solvent but are sufficiently heavy to favor inertial effects.

When studying stochastic equations, the most powerful tool available is the Fokker–Planck equation. This equation describes the evolution of the probability distribution of x and p: $\mathrm{P}(x, p, t)$, where p is the momentum of x. It depends on the initial distribution $\mathrm{P}(x, p, 0)$. In particular, if we aim to perform kinetic estimations, the Fokker–Planck equation can provide the average time it takes to reach a target state from a given initial state, which is precisely what is needed.

However, in the case of the underdamped regime, only one solution of the Fokker–Planck equation is known, which corresponds to the scenario where the potential is flat [93]. Because the potential is crucial for chemical reactions, this solution cannot be used. However, there is a stochastic regime for which the Fokker–Planck equation has a solution: the overdamped Langevin dynamics.

3.2.2 Overdamped model

When friction increases, another type of Langevin equation can be established by neglecting the inertial term $m\ddot{x}$. This is the overdamped equation, introduced here as a limit case of the underdamped:

$$\dot{x} = \frac{f(x)}{\gamma * m} + \frac{1}{\gamma * m}\xi \tag{3.2.6}$$

$$\dot{x} = -\beta D f(x) + \sqrt{2D} W \tag{3.2.7}$$

where D is the diffusion constant: $D = \frac{1}{\beta \gamma m}$.

In this case, the stochastic equation is first order. In this context, the Fokker-Planck equation becomes the Smoluchowski diffusion equation and has the following expression [94, 95]:

$$\frac{\partial}{\partial t}P(x,t|x_0,t_0) = \frac{\partial}{\partial x}D\left(\frac{\partial}{\partial x} - \beta f(x)\right)P(x,t|x_0,t_0) \tag{3.2.8}$$

where x_0 and t_0 represent the initial position and the origin of time, respectively. This expression must be solved for different forces f [95]. In particular, the solutions of the Smoluchowski equation serve as a theoretical basis for most models used to predict the kinetics of chemical processes.

3.2.3 Generalized model

When the time scale of resolution of the model goes below the characteristic time of fluctuations of the solvent, the overdamped and underdamped equations become insufficient in describing the system behavior. In this case, the friction response comes with a decay, representing the characteristic response time of the solvent fluctuations. To capture this "decay", we use a history of the velocities. It appears as if the preceding values of the particle velocity (the degree of freedom followed) are stored in the behavior of the solvent, indicating that the solvent has a temporary memory. The function that translates this memory is the memory kernel $\mathrm{K}(\tau)$, the critical component of Generalized Langevin Equation (GLE):

$$m\ddot{x}(t) = f(x(t)) - \int_{-\infty}^{t} \mathrm{K}(\tau)\dot{x}(t-\tau)\,d\tau + \xi(t) \tag{3.2.9}$$

When integrating this equation, the new value of $x(t)$ depends not only on the last preceding instants, but on all preceding values of the velocity. Thus, the dynamics becomes non-Markovian. It should be noted that when $\mathrm{K}(\tau)$ is a Dirac delta function, we revert to the underdamped model.

A discussion of the generalization of the fluctuation-dissipation theorem in this type of dynamics is still open [91]. However, in this work, we will assume that this theorem still holds. In the generalized case, it becomes:

$$\langle \xi(0)\xi(\tau)\rangle = \beta^{-1}\mathrm{K}(\tau) \tag{3.2.10}$$

Here, the noise is autocorrelated, meaning each new value of the noise depends on its history, becoming a "colored noise" [96, 97].

As GLE is an equation from which both the underdamped and overdamped models can be derived, it can represent a wide variety of processes. However, the associated Fokker-Planck equation remains unsolved.

The use of models with overdamped, underdamped, or generalized models to study chemical reactions in solution is still an area of active research [98, 96, 99, 100, 89]. The

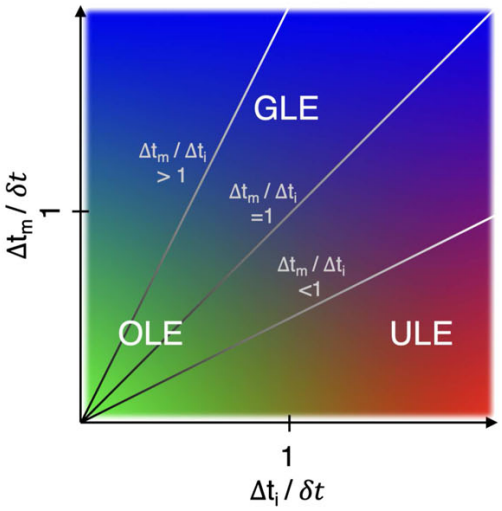


Figure 3.2.1: The diagram of the application domains of the generalized, underdamped, and overdamped Langevin models. δt is the time resolution of the model, $\Delta t_i = \gamma^{-1}$ is the time necessary to reach thermal equilibration, and Δt_m is the decay time of the memory kernel (*i.e.*, the fluctuation time of the solvent).

main question is whether the generalized model is necessary to efficiently represent the dynamics of a chemical reaction or if the overdamped model is sufficient. The latter would be tempting to use because it allows for the application of the Smoluchowski equation. However, the estimated rate with overdamped models is often several orders of magnitude away from experimental values [101, 102]. A summary of the application domain of these three models is illustrated in Figure 3.2.1, taken from Girardier's article [89].

To explain this discrepancy, we will review the available methods for calculating reaction rates from a theoretical perspective.

3.3 Rate estimations

The calculation of the kinetics of a chemical reaction is at the intersection of three fields: pure mathematics, statistical physics, and theoretical chemistry. To illustrate a first layer of chemical kinetics, we introduce a fictitious reaction:

$$A + B \rightleftharpoons C$$

In this case, the two possible reactions are the forward reaction, indicated as **1**, and the backward reaction, denoted as **-1**.

The "speed" of each reaction at high dilution, which is the number of occurrences of the reaction in a certain fixed amount of time, is noted as follows:

$$v_1 = k_1 \frac{[\mathrm{A}]}{c_0} \frac{[\mathrm{B}]}{c_0} \tag{3.3.1}$$

$$v_{-1} = k_{-1} \frac{[\mathrm{C}]}{c_0} \tag{3.3.2}$$

where [A], [B], and [C] are the molar concentrations and c_0 is a constant standard concentration, usually $1\mathrm{mol} \cdot L^{-1}$. k_1 and k_{-1} are the rate constants, which are the parameters of the system.

The time evolution of the system relies on differential equations:

$$\frac{d[A]/c_0}{dt} = \frac{d[B]/c_0}{dt} = v_{-1} - v_1 \tag{3.3.3}$$

$$\frac{d[C]/c_0}{dt} = v_1 - v_{-1} \tag{3.3.4}$$

These equations are first-order differential equations that can be solved in the general case. From this, the only remaining difficulty is to determine the rate constants.

3.3.1 The Mean First Passage Time (MFPT)

The straightforward way to estimate the rate is to measure the average time it takes for a dynamic model to cross the barrier, known as Mean First Passage Time (MFPT). This is the "brute force" estimation of the rate. If we launch a large number of dynamic simulations at the bottom of the reactant well and for each launch record the time it takes to reach the bottom of the product well, t_i, then:

$$k^{-1} = \langle t_i \rangle \tag{3.3.5}$$

With an infinite amount of time and computational power, this method could be achieved by sampling a Langevin model or even by performing AIMD long enough. This method has the advantage of being independent of any RC: the mechanism of the transition is not taken into account in the estimation.

Of course, the curse of time prevents us from obtaining these results, even with the cheapest Langevin model, considering the barrier we have to cross. However, in the overdamped case, this value can be calculated directly owing to the Smoluchowski equation by integrating the free-energy landscape:

$$k^{-1} = \int_{x_0}^{b} \frac{e^{\beta F(x)}}{D} \left(\int_{a}^{x} e^{-\beta F(x')} \, dx' \right) dx \tag{3.3.6}$$

where $F(x)$ is the free energy curve associated with x, x_0 is the initial position of the dynamics, assimilated to the bottom of the reactant well, a is the position of a reflective boundary located in the opposite direction of the barrier, and b is an absorbing boundary located at the bottom of the product well.

An approximation of these rates, also based on the overdamped model, was introduced by Kramers in 1940 in the case of a diffusive regime [93, 101]:

$$k = \frac{\omega_A \omega_{TS}}{2\pi\gamma} e^{-\beta F^\ddagger} \tag{3.3.7}$$

where ω_A and ω_{TS} are the frequencies of the harmonic approximation of the potential for the well and the barrier, respectively.

3.3.2 The transition state theory

The transition state theory is based on the assumption that the event of crossing the barrier is the combination of two independent events: reaching the barrier top and having the right orientation of velocity at the top. Part form this a straightforward measure of the rate take the form:

$$k^{-1} = \frac{1}{\beta h} e^{-\beta F^\ddagger} \tag{3.3.8}$$

where $F^\ddagger$ is the activation energy and h is the Planck constant. [103]

This method neglects an important aspect of rate estimation, which is the recrossing effect: the phenomenon in which trajectories that reach the top of the barrier can recross back almost immediately due to fluctuations in the solvent that provide the appropriate momentum. To account for this, we can use the reactive flux method.

3.3.3 The reactive flux

The reactive flux method estimates the rate from transition path sampling data. Recrossing effects are represented when some trajectories exceed the barrier and then recross as a result of the influence of the solvent. This method depends on the time [104, 105, 106]:

$$k(t) = \frac{\langle \dot{x}(0)\delta[x(0) - x^*] h_P(x(t)) \rangle}{\langle h_R(x(t)) \rangle} \tag{3.3.9}$$

where $h_R(x)$ and $h_P(x)$ are the indicator functions of the wells of the reactants and products, respectively, and can be defined using a OP. Here, x^* is the value of x at the top of the barrier height, referenced to the separatrix. Since spontaneous transitions are rare events, $k(t)$ reaches a plateau over a wide range of t values that are higher than the typical descent time (*, that is,*, the time required to descend the barrier) but lower than MFPT.

By sampling N trajectories from the top of the barrier (that is, at x^*), we can numerically estimate the rate as the product of the time-dependent average over the trajectories conditioned on $h_P(x(t))$ and an exponential term that depends on the free-energy barrier:

$$k(t) = \frac{\sum_{j=1}^{N} \dot{x}_j(0) h_P(x_j(t))}{N} \frac{e^{-\beta F^\ddagger}}{\int_{\Omega_R} e^{-\beta F(x)}\, dx} \tag{3.3.10}$$

where Ω_R is the region of the reactants.

This numerical estimation is based on the work of Palacio-Rodriguez [106] and can be performed using AIMD enhanced sampling simulations: TPS for averaging and US to determine the free energy curve along x.

3.4 Conclusion of this chapter

In this chapter, we presented the statistical tools that are funded for the kinetics of chemical processes, specifically the Langevin models and the Fokker-Planck equation. We also discussed the numerical tools used to estimate kinetics, which are closely linked with Langevin models, particularly the solvable overdamped model. Despite advances, a significant portion of the error in kinetic estimations still arises from incorrect estimations of the barrier height.

In the realm of chemical reactions in solution, a substantial part of the barrier is crossed during the cleavage and formation of covalent bonds. This indicates that, at this moment, the impact of the solvent on the dynamics is minimal, the friction is low, and the inertial term becomes significantly important. Consequently, the use of underdamped or generalized Langevin models may be appropriate near the top of the barrier.

However, in wells, the behavior of the system is different. For solute compounds in solution or intramolecular reactions that involve exploring various conformers to achieve a reactive state, the dynamics is more significantly influenced by friction with the solvent. This suggests that a suitable model to accurately capture the full dynamics of a chemical reaction might include a position-dependent friction coefficient or a memory kernel [96]. The inference of this type of model for chemical processes is challenging and is still under investigation and will be discussed in detail in the results section.

This chapter concludes the methodology part of this thesis in which we introduce all tools used for studying chemical reactions in solution, from molecular dynamics and enhanced sampling techniques to kinetic inference.

Part II

Applications

Chapter 4

Application to a S_N2 reaction in solution

4.1 Presentation of the article

This article is the result of work initially performed by a former Ph.D. student of our research group, Théo Magrino. As he made a significant part of the calculations and writing, he remains the first author. During my first year, part of my work involved familiarizing myself with the tools he developed and verifying the reliability of the article. Some of the proposed results have been reproduced for reliability, and some have been updated. Since the article had unresolved issues that prevented its publication, we worked to complete the missing information and reformat a significant portion of the article. We also changed some of the notation and formulas for better readability.

This article presents a complete workflow for the study of chemical reactions in solution, from unbiased AIMD to the Umbrella Sampling simulation process. The article emphasizes two main aspects of this protocol: the efficiency of the transition path sampling process to sample the reaction pathway and reach the MFEP and the quality of the generated data-driven PCV. This article is a milestone as it serves as a first test of the quality of our protocol and an exploratory article in which we examine the limits and possibilities of the tools we used and experiment with new ones.

In the initial version, this article was also supposed to test the feasibility of using replica exchange umbrella sampling, but the results this method produced for our system led us to include this study in the SI for now and to develop further work focused on this powerful tool in future studies.

This article has already been published in J. Phys. Chem. A. The supporting information associated with it is in Appendix B.

Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry

Théo Magrino,[†,‡] Léon Huet,[†,‡] A. Marco Saitta,[†] and Fabio Pietrucci[*,†]

†*Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS UMR 7590, Paris 75005 FRANCE*

‡*These two authors contributed equally*

E-mail: fabio.pietrucci@sorbonne-universite.fr

Abstract

Reaction coordinates are an essential ingredient of theoretical studies of rare events in chemistry and physics, since they carry information about reaction mechanism and allow the computation of free-energy landscapes and kinetic rates. We present a critical assessment of the merits and disadvantages of heuristic reaction coordinates, largely employed today, with respect to coordinates optimized on the basis of reliable transition path sampling data. We take as test bed multi-nanosecond *ab* initio molecular dynamics simulations of chloride S_N2 substitution on methyl chloride in explicit water. The computational protocol we devise allows the unsupervised optimization of agnostic coordinates able to account for solute and solvent contributions, yielding a free energy reconstruction of quality comparable to the best heuristic coordinates, without requiring chemical intuition.

Introduction

Molecular dynamics (MD) simulations represent a powerful tool for the study of chemical reactions, able to account for temperature, pressure, entropy and solvent effects. The two main limitations of this methodology are the limited accuracy of the computationally affordable approximations of quantum-mechanical interatomic forces, and the "rare event problem", i.e., the wide gap between computational and experimental timescales. MD trajectories based on density functional theory (DFT) reach today the nanosecond timescale for systems including up to about one thousand atoms.[1]

To overcome this second limitation, two main classes of techniques have been developed, namely transition path sampling and biased simulations.[2] Especially in the latter class, a crucial step consists in the identification of collective variables (CVs), also called reaction coordinates (RCs), able to track the detailed mechanism of the transformation: the accuracy of the predicted barriers and rates strongly depend on the quality of the RC. More generally, besides the interest in the context of simulations, the RC is an important tool for getting human insight into the relevant chemical reaction features.

In principle, a single RC is sufficient to describe (and to accelerating the sampling of) a chemical reaction connecting two metastable states via a single mechanism: the optimal RC is customarily identified with the committor function $\phi(x)$, i.e., the probability to reach products before reactants when evolving from a given atomic configuration with initial equilibrium velocities.[3–6] In practice, systems com-

prising hundreds of molecules pose a challenge to the identification of such optimal coordinate: for instance, experimental evidence indicates that the solvent is able to play an important role – nontrivial to rationalize – in defining the reaction mechanisms.[7] On the other hand, a large number of heuristic CVs could be defined based on chemical intuition.

An interesting class of reactions in the context of the present discussion is represented by halides nucleophilic substitution on alkyl halides, in particular the single-act S_N2 associative bimolecular mechanism, one of the first reactions receiving detailed attention in modern chemistry.[8,9] In this class, chloride S_N2 substitution on methyl chloride, $CH_3Cl + Cl^- \rightleftarrows CH_3Cl + Cl^-$ has been the object of experimental work in gas phase and in solution, especially in water.[10–14]

In this reaction, reactant and product states are equivalent, implying a zero free-energy difference between them. The inability to distinguish reactants from products poses a problem to direct kinetic measurements: the activation free energy in water has been estimated at 300 K via empirical extrapolation of the measured barriers of similar non-symmetric reactions as 26.5 Kcal/mol in Ref.,[15] from Marcus theory[16] with a linear formulation in terms of known thermodynamic constants, and 26.6 kcal/mol in Ref.[17] from Swain-Scott equation[18] (no error bars reported). We will refer to this value as "experimentally-derived".[19] Forward and backward rate values have been on the other hand measured for the asymmetric S_N2 reaction $CH_3Br + Cl^- \rightleftarrows CH_3Cl + Br^-$, yielding $k_1 = 4.13 \pm 0.22$ and $k_{-1} = 0.93 \pm 0.06$ $(10^{-7}s^{-1})$, respectively.[20]

Considering its relative simplicity, methyl chloride-chloride S_N2 substitution has therefore been a guinea pig of theoretical chemistry since early works on frontier orbitals by Kenji Fukui, Nobel price in chemistry.[21] The reaction has been the object of early application of different techniques, from variational transition state theory[22] to Grote-Hynes theory,[23–26] as well as metadynamics more recently.[27]

Solvent effects can be intuitively expected to be significant in this reaction, due to the redistribution of electronic charge taking place along the transition. Indeed, experiments showed both qualitative and quantitative changes between gas-phase and solution free energy profiles.[15,28–30] In solution, the presence of the solvent makes the reactant/product state, rather than just one meta stable state, but a composition of multiple meta-stable states issued of the different number of intercalated water molecules, and one stable state that correspond to the fully screened Cl^-/CH_3Cl attraction. Second, the transition state – less polar than reactants and products – is destabilized in water solution, due to weaker solute-solvent bonding,[11,12,15,31,32] consistently with the trend of decreasing activation barrier with decreasing solvent polarity.[15]

Notwithstanding their important role, it is unclear how to include solvent degrees of freedom into the RC. As a result, for convenience reasons most simulation studies traditionally employed a simple RC excluding altogether the solvent variables, in the form of the difference between the two carbon-chlorine distances (thus capturing the evident symmetry of the reaction), denoted $d_1 - d_2$ in our work. This heuristic, hence suboptimal RC definition is a potential source of hysteresis in enhanced sampling simulations, i.e., of insufficient equilibration of the solvent degrees of freedom on the short timescale of the simulation.[33] For instance, in Langevin models of reactions based on the Grote-Hynes theory,[23,24] it has been shown that the solvent non-instantaneous rearrangement to solute transformation leads to significant memory effects in the friction and noise forces of the projected dynamics. Clearly, the assumption of timescale separation between the solvent and solute degrees of freedom inherent in the choice of $d_1 - d_2$ as RC is not devoid of risks.

Reports in the literature indicate two issues that render challenging the computational study of the methyl chloride-chloride S_N2 reaction in water. The first issue concerns the limited accuracy of interatomic force calculations, the second issue the limited statistical precision of free-energy calculations (that rely on the ergodic behavior of many-body MD trajectories,

a feature difficult to assess). Clearly, in presence of unsatisfactory results it can be arduous to disentangle the contributions of the two problems.

Since the 1980's, several computational studies were able to reproduce with good accuracy the experimentally-derived free-energy barrier of 26.6 kcal/mol at 300 K[15,17] by adopting atomistic force fields[28] or a combination of force fields and quantum mechanical calculations (QM/MM).[34–36] However recent studies pointed out the sensitivity of barrier values to the chosen QM/MM method,[37] ranging from 38 kcal/mol[37] to 14-15 kcal/mol,[38] respectively. For instance, Ref.[37] emphasizes the strong influence of the classical charge assignment on the transition state energy. When computationally feasible, like in the case of the reaction under scrutiny, a fully quantum description of the system is expected to be more robust, provided system size-effects are avoided with the use of sufficiently large simulation boxes.

As stated above, the precision and accuracy of computational results (mechanisms, free energies and rates) depends critically also on the enhanced sampling protocol, which in turn relies upon the choice of RCs for accelerating the dynamics and analyzing the results.[2,39] For instance, an early fully-DFT study pointed out solvent-induced hysteresis effects (as large as 5 kcal/mol) on the calculation of free-energy profiles with thermodynamic integration when using solvent-less RCs and few-picoseconds trajectories.[33] Two approaches to avoid such artifacts would be the use of longer simulations (ensuring proper equilibration), without knowing how much exceeding time would be necessary, and the use of optimized RCs,[27,40] that should in principle explicitly include solvent degrees of freedom. The latter path has been recently explored via likelihood optimization of RCs starting from transition path sampling data, albeit only at the level of analysis and without testing the candidate RC with the application of bias to accelerate and sample the reaction.[38]

In the present work, we propose a protocol aimed at exploring reaction pathways and reconstruct free energy landscapes for chemical reactions in solution without employing any educated guess about the mechanism or the RC. We believe this to be an important task, since a broad application of MD simulations to obtain structural, thermodynamic and kinetic insight on a wide range of reactions will remain unreachable until reliable unsupervised approaches will be demonstrated.

Taking the paradigmatic case of the methyl chloride-chloride S_N2 reaction in water as benchmark, we show that it is possible to start from the sole knowledge of reactants and (putative) products and achieve detailed quantitative insight on the reaction by applying fully ab initio MD and a battery of state-of-the-art enhanced sampling methods. A preliminary agnostic exploration of reaction mechanisms is performed with metadynamics combined with flexible CVs. The latter track changes of coordination pattern between reactants and products,[41] allowing to discover favorable transition pathways without making any hypothesis on them. This kind of CVs can address a wide variety of chemical reactions in solution on the same footing, as showed in a recent study of the Strecker synthesis of glycine,[42] a complex multi-step and multi-component chemical reaction. Next, solid and reliable transition path sampling techniques allow refining in a fully unbiased way the mechanism found by metadynamics. At this stage, the pathways can be confidently analyzed to identify a RC able to track the detailed structural changes associated with the transition, and such RC is finally employed to reconstruct a free energy profile with umbrella sampling and weighted histogram analysis.

Our approach is data-driven and removes at the best human subjective biases from the study of chemical reactions. The pertinence and effectiveness of the protocol is tested in a twofold way: by assessing the quality of the predicted RC via machine learning techniques (i.e., principal component analysis and likelihood maximization of the committor description[4]), and by comparing with traditional RC proposals based on chemist intuition or trial and error. Merits and disadvantages of the two philosophies, data-driven versus heuristic-driven, are critically discussed.

Theoretical Methods

Our protocol for the study of chemical reactions employs finite-temperature *ab initio* molecular dynamics simulations, and consists of two complementary steps, addressing the exploration of mechanisms and their statistical sampling:

- Mechanism exploration: transition pathways are agnostically explored using metadynamics (MTD) based on low-resolution two-state path CVs that require only specification of reactants and putative products; the reaction mechanism is further refined with transition path sampling (TPS), to construct precise multi-state path CVs.
- Free energy calculation: efficient umbrella sampling (US) simulations based on the latter CVs are used to compute precise free-energy landscapes via weighted histogram analysis.

The protocol is schematically represented in figure 1, the first part with blue blocks and the second part with green blocks.

Molecular dynamics simulations

All molecular dynamics simulations were performed at the DFT level with the Perdew–Burke–Ernzerhof functional,[43] as implemented in the CPMD code.[44] We employed a periodically repeated cubic box of 14.48 Å with an atomic composition corresponding to $CH_3Cl + Cl^- + K^+ + 98\,H_2O$. The resulting density is 1.034 kg/L, with a concentration of solutes of 0.547 mol/L. We used Martins-Troullier pseudopotentials,[45] a plane-wave expansion of Kohn-Sham orbitals up to a cutoff of 80 Ry and an orbital optimization convergence of 10^{-5}. Born-Oppenheimer molecular dynamics simulations were performed with a time step of 0.48 fs, in the canonical ensemble (NVT) at 300 K based on the Nosé-Hoover thermostat for each ionic degree of freedom with a 3000 cm^{-1} frequency and chain length equal to 4. Hydrogen atomic mass was set to 2 a.u. to increase numerical stability.

We generated a total of ∼100 ps MTD trajectories, ∼1600 ps TPS trajectories and ∼1200 ps US trajectories. All input details are freely available on PLUMED-NEST (www.plumed-nest.org), the public repository of the PLUMED consortium.[46]

Topology-based path collective variables

Enhanced sampling simulations were performed using a modified version of the plugin Plumed 1.3.[47] A guide to the use of the CVs employed in this work, as well as source code and example input files for Plumed 1.3 and Plumed 2.x can be freely downloaded from https://sites.google.com/site/fabiopietrucci and from PLUMED-NEST (www.plumed-nest.org,[46]).

Coordination pattern-based path CVs[41,48] are used as an effective compromise between the need of a high-dimensional space of coordination numbers to discriminate structures, and the need of a low-dimensional projection to sample efficiently the free-energy profile.[49] The CVs are defined starting from reference atomic structures X_α, with $\alpha \in [\![1, N]\!]$, tracking the progress of a chemical reaction. We employed $N = 2$ for low-resolution CVs in the explorative part and $N = 12$ for high-resolution CVs in the quantitative part of the protocol.

Indicating with $x(t)$ the MD configuration at a given time, path CVs are defined as :

$$\begin{cases} s(t) & = \dfrac{\sum_{\alpha=1}^{N} \alpha \times \exp(-\lambda \mathrm{D}[x(t), \mathrm{X}_\alpha])}{\sum_{\alpha=1}^{N} \exp(-\lambda \mathrm{D}[x(t), \mathrm{X}_\alpha])} \\ z(t) & = \dfrac{-1}{\lambda} \log \sum_{\alpha=1}^{N} \exp(-\lambda \mathrm{D}[x(t), \mathrm{X}_\alpha]) \end{cases} \tag{1}$$

where $\mathrm{D}[x(t), \mathrm{X}_\alpha]$ is a distance metric based on patterns of coordination numbers, as introduced in Ref.:[41]

The coordination function $c_{i\sigma}$ approximately counts the number of atoms of a set σ (here, each of the available chemical elements) connected to a given atom i (here carbon and chlorine atoms), and is expressed based on inter-

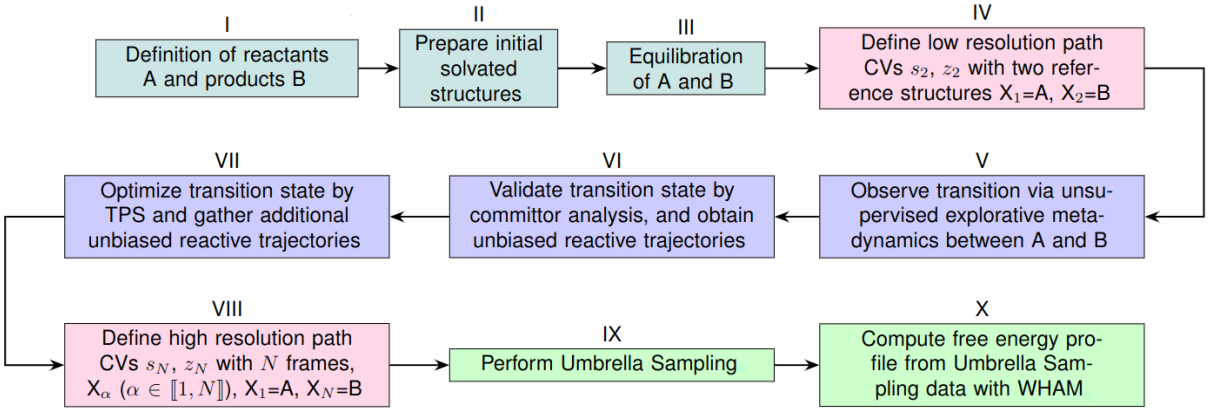


Figure 1: Schematic algorithm depicting our simulation protocol. Dark green blocks (I, II, III) indicate initialization of the protocol. Blue blocks (V, VI, VII) indicate agnostic explorative steps. Light green blocks (IX, X) indicate quantitative sampling steps. Pink blocks (IV, VIII) indicate pivotal steps were path CVs are defined.

atomic distances d_{ij} :

$$\mathrm{c}_{i\sigma} = \sum_{j\in\sigma} \frac{1-\left(\frac{d_{ij}}{d_0}\right)^8}{1-\left(\frac{d_{ij}}{d_0}\right)^{14}} \qquad (2)$$

i.e., a sum over switching functions monotonically increasing from 0 to 1 for decreasing distance d_{ij} The parameter d_0 was set to 2.2 Å for Cl-Cl, Cl-C, Cl-O, Cl-K and C-K pairs, 1.8 Å for C-O and Cl-H pairs and to 1.5 Å for C-H pairs.

As an illustration, reference coordination matrices $\mathrm{c}_{i\sigma}(\mathrm{X}_\alpha)$ corresponding to reactants an products of two-state path CVs are presented in figure 2. In the case of locally-stable states, matrices used as references correspond to average coordination values over equilibrium MD trajectories.

For the case $N = 12$, $\mathrm{D}[\mathrm{X}_\alpha, \mathrm{X}_{\alpha+1}]$ is approximately equal for each α (see below for the selection of the reference structures). Both for $N = 2$ and for $N = 12$, the parameter λ is set from the relation $\exp(-\lambda\,\mathrm{D}[\mathrm{X}_\alpha, \mathrm{X}_{\alpha+1}]) \approx 0.1$, that ensures relatively smooth free energy landscapes and a good resolution of different atomic configurations.[48] For comparison, when $\lambda \to \infty$, s, z become discontinuous, assuming the value $s = \alpha$ of the closest reference structure, with $z = \mathrm{D}[x(t), \mathrm{X}_\alpha]$; when $\lambda \to 0$, the variables become unable to resolve different structures. Different abbreviations for the most important CVs in these work are summarized in table 1.

Table 1: Abbreviations used for CVs in this work.

d_i	C-Cl_i distance
$c(i, \sigma)$	coordination number of atom i with respect to atoms of set σ
c_i	c(Cl_i,C) coordination number
s_2, z_2	low-resolution two-state path CVs
$s_{12,\mathrm{MCA}}$, $z_{12,\mathrm{MCA}}$	high-resolution 12-state path CVs built from single committor analysis trajectory issued from MTD
$s_{12,\mathrm{TPS}}$, $z_{12,\mathrm{TPS}}$	high-resolution 12-state path CVs built from TPS trajectories

Metadynamics exploration of reaction mechanisms

Metadynamics is adopted as an efficient approach to accelerate the escape from the free-energy minimum corresponding to reactants until reaching the products minimum. In this technique, a history-dependent external bias is added to the interatomic potential, in the form of sum of Gaussian functions localized along the trajectory in CV space.[50,51] This allows exploring a putative reaction mechanism and transition state region, that are expected to be close

		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 2: Reference coordination matrices of two-state s_2 and z_2 path CVs employed in metadynamics simulations. Values are averages over reactants equilibration, with Cl_1 and Cl_2 simply switched between reactants and products.

to the optimal ones – i.e., those sampled with high probability in a hypothetical unbiased simulation of sufficient duration – under the hypothesis of slow bias deposition along the optimal RC.

In practice, the high computational cost of ab initio MD limits the slowness of bias deposition, while our choice to avoid any educated guess on the transition mechanism – an output and not an input of our protocol – excludes the possibility of knowing the optimal RC beforehand. As a compromise, we employ a flexible definition of path CVs that, albeit lacking any information about the pathway, was proven in previous works to track efficiently a range of chemical reactions in water. The identification of the optimal RC is performed in a later step of the protocol, after TPS refinement of the mechanism.

We adopted low-resolution two-state path CVs s_2 and z_2 built from the equilibrated reactants and products as reference states (step IV in figure 1). The flexibility of the second coordinate allows exploration of different transition states, i.e., different reaction mechanisms, within the same simulation. We have shown in ref.[41] and[42] that this technique can result in the discovery of intermediate metastable states that have not been specified in the path CVs definition. The Gaussian parameters are $\sigma_{s_2} = 0.03$, $\sigma_{z_2} = 0.05$, with a height of 3.14 kcal/mol (0.005 au) and a deposition interval of 50 MD steps (24 fs).

To limit the exploration within the relevant region of configuration space, we applied a semi-parabolic potential restraining the distance between the carbon atom and the entering chlorine when exceeding 7.5 Å, with force constant equal to 1.8 kcal.mol^{-1}.Å^{-2}. This limit corresponds approximately to three solvation shells (defined as local minima in chlorine-oxygen radial distribution function), and 59 integrated water molecules in the corresponding radius around chlorine.

The MTD simulation is stopped after observing the first reactive transition, in order to provide candidate transition state structures without attempting a quantitative estimation of the free energy landscape.

Committor analysis and transition path sampling

Starting from the first putative transition pathways obtained by MTD, transition state configurations are identified with committor analysis. A set of 20 atomic configurations are extracted from the barrier region (i.e., the CV-space region between the reactants and products minima), and from each configuration 20 to 40 unbiased MD trajectories are generated with initial velocities randomly drawn from the Maxwell-Boltzmann distribution at $T = 300$ K. The fraction of trajectories relaxing into products is an estimate of the committor function, and a value close to 0.5 indicates a member of the transition state ensemble.

Within the latter ensemble, in principle, structures can have different Boltzmann probabilities: in general, the metadynamics bias applied along a non-optimal RC does not guarantee that the most probable transition state structures are systematically sampled. However, for a precise identification of the re-

action mechanism and of the optimal RC it is necessary to characterize the most probable transition state configurations as well as the most probable reactive pathways passing through them. For this task, techniques belonging to the TPS family are the golden standard.[52,53] Such techniques iteratively sample members of the transition state ensemble, relaxing towards the members that have the highest Boltzmann probability: starting from an initial phase-space point, a new point is generated according to some procedure, and short unbiased trajectories are shot forward and backward in time, inspecting if the resulting path connects reactants to products.

For TPS initialization we employ two committor analysis trajectories starting from the same MTD point, one committed to reactants, the other committed to products, joined to form a single reactive path (inverting the time direction in one of them), setting the time origin at the shooting point. This initial path is indicated as $X \equiv \{x(-t_i)..x(0)..x(+t_f)\}$. We tested two different TPS algorithms:

(i) Aimless shooting (AS).[54] We employ here the two-point flexible length version of aimless shooting, proposed in Refs.[55,56] Denoting Δt the time step parameter the algorithm proceeds as follows:

1. Select $x(-\Delta t)$ or $x(+\Delta t)$ with probability 1/2 from the last accepted trajectory as new shooting point.

2. Draw velocities from the Maxwell-Boltzmann distribution (at 300 K) and propagate two trajectories, one forward in time, one backward, until they reach reactants or products.

3. If one of the trajectories reached reactants and the other products, accept the shooting point and the two trajectories, joining them into a new reactive path X setting $t = 0$ at the shooting point and inverting time for the backward part.

4. Iterate.

(ii) Shooting from the top (SFT).[57] After selecting s_2 as order parameter and a given interval $[s_A, s_B]$ as the barrier-top region, the algorithm proceeds in the same way as the previous one, with only a modification in the first step of each iteration:

1. Select a point $x(t) \in X$ such that $s_2(x(t)) \in [s_A, s_B]$, with uniform probability, as new shooting point.

In both algorithms, products are defined by $s_2 > 1.8$, and reactants by $s_2 < 1.2$. In SFT TPS, we slightly modified the acceptance rule compared to ref.,[57] as critically discussed in supplementary materials.

We performed AS testing different Δt=2.5, 10 and 15 fs (denoted $AS_{2.5}$, AS_{10} and AS_{15}, respectively). We note that both too short as well as too long Δt are expected to reduce AS efficiency, due to slow diffusion in configuration space or to low acceptance of new configurations, respectively (see the Results section).

We performed SFT with different s_2 ranges, respectively [1.30,1.70], [1.35,1.65] and [1.45, 1.55] (denoted $SFT_{1.30-1.70}$, $SFT_{1.35-1.65}$ and $SFT_{1.45-1.55}$, respectively). $SFT_{1.35-1.65}$ was retained for further analysis, and has been performed twice, because of large uncertainties in likelihood estimates (see below).

For each TPS algorithm and each parameter choice, $\sim$ 300 shooting steps have been performed, each one of typical duration 2×0.15 ps. For the two $SFT_{1.35-1.65}$ simulations, referred to as set A and B, 520 shootings have been performed. Approximately $\sim$ 200 shootings were accepted for each set.

Definition of high-resolution multi-state path CVs

The choice of the number of reference structures in the definition of path CVs leaves the flexibility to specify a reference pathway with lower or higher resolution. The initial metadynamics exploration (based on two-state path CVs s_2 and z_2) as well as the committor analysis and TPS simulations provide detailed information about the reaction mechanism. Before proceeding to the extensive sampling of free-energy landscapes (step IX figure 1), this information is included in the definition of high-resolution

multi-state path CVs, using 12 references coming from either metadynamics-based committor analysis (denoted $s_{12,\mathrm{MCA}}$ and $z_{12,\mathrm{MCA}}$), or from TPS $\mathrm{SFT}_{1.35-1.65}$ trajectories (denoted $s_{12,\mathrm{TPS}}$ and $z_{12,\mathrm{TPS}}$). This corresponds to step VIII in figure 1.

$N = 10$ reference structures are extracted from a selected transition pathway, employing the algorithm and code proposed in Ref.,[42] with a minor modification, as described here. First, $N-2$ structures ($k = 2, ..., N-1$) are randomly chosen and ordered according to s_2. The first ($k = 1$) and last ($k = N$) points, are chosen as the coordination patterns of reactants and products, respectively, averaged over equilibration trajectories. A fictitious elastic energy is associated to the $N-1$ distances among consecutive references $\mathrm{D}_{k,k+1}$ (see equation **??**, $k \in [\![1, \mathrm{N}-1]\!]$), with the rest length of each segment of the elastic band equal to the average length of all segments, while a second contribution accounts for angles θ_k formed between consecutive path segments:

$$E = \sum_{k=1}^{\mathrm{N}-1} \left(\mathrm{D}_{k,k+1} - \frac{1}{\mathrm{N}-1} \sum_{l=1}^{\mathrm{N}-1} \mathrm{D}_{l,l+1} \right)^2 + \beta \sum_{k=2}^{N-1} [\max(\theta_k - \theta_{thresh}, 0)]^2 \quad (3)$$

The first term favors approximately equidistant points in the space of coordination patterns, whereas the second term tends to reduce the length of the chain by favoring a low curvature. The parameter β governs the relative importance of the angular part. A Monte-Carlo optimization of the energy function is performed, with moves consisting in selecting a new point for reference k by choosing a random structure between references $k-1$ and $k+1$. The acceptance criterion was standard Metropolis acceptance criterion, using a variable temperature cooling down as the energy decreases, defined as $T = \alpha E_{last}$, where E_{last} is the energy of the last accepted configuration. For $\mathrm{s}_{12,\mathrm{MCA}}$ we used $\alpha = 0.45$, $\beta = 0.45$, and θ_{thresh} =0°. For $\mathrm{s}_{12,\mathrm{TPS}}$ we used $\alpha = 0.5$, $\beta = 0.6$, and θ_{thresh} =36°. Typically about 500 000 iterations are sufficient to observe the stabilization of the reference set. We remark that convergence is greatly improved, by allowing the random swap of k, k+1. Compared to the algorithm in Ref.,[42] introduction of the parameter θ_{thresh} also improved convergence. The algorithm is tolerant to variations in the parameters α, β and θ_{thresh}, reliably finding a set of references where all side by side distances are equivalent and all angles are lower than θ_{thresh}.

Eventually, two artificial reference patterns are added before the first and after the last reference (leading to $N = 12$), by linearly extrapolating the first and last chain segments, to avoid metastable states to appear as spikes in the free-energy landscape.

Figure 3 represents the selected reference structures in the (c_1,c_2) plane.

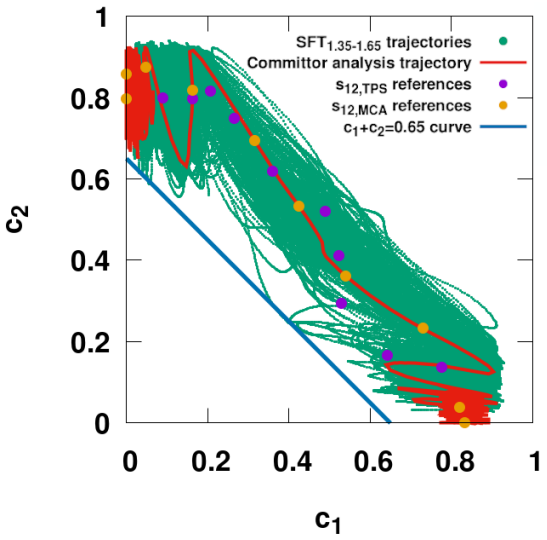


Figure 3: Representation of reactive trajectories in the (c_1,c_2) plan: the red curve represents a pair of committor trajectories shot from metadynamics, while the green curves represent a set of transition path sampling trajectories (referred to as $\mathrm{SFT}_{1.35-1.65}$ in the text). Orange and violet points represent the reference configurations selected by the algorithm to build path CVs $\mathrm{s}/\mathrm{z}_{12,\mathrm{MCA}}$ and $\mathrm{s}/\mathrm{z}_{12,\mathrm{TPS}}$, respectively (see Methods section). The blue line emphasizes that $c_1 + c_2$ never falls below 0.65, due to the concerted nature of the mechanism.

Transition path sampling-based optimization of the reaction coordinate

TPS trajectories contain rich and unbiased information about the mechanism and the kinetics of the reaction: besides being used to define multi-state path CVs, they were employed as input data for machine learning techniques aimed at identifying the optimal RC in a given subspace of configuration space.

First, we performed principal component analysis (PCA) of accepted reactive trajectories harvested from SFT TPS with $s_2 \in [1.35,1.65]$ (data sets A and B), using as variables the set of coordination numbers in figure 2. We employed the standard python library scikit-learn,[58] with full svd solver. Data were centered but not normalized, as all variables are coordination numbers, adimensional and with the same physical meaning.

Second, we analyzed the same trajectories (limited to data set A) with the non-inertial likelihood optimization scheme for the RC..[4,54,55] The likelihood is evaluated for TPS shooting points: each configuration corresponds to specific values of a given set of CVs $\{\mathrm{CV}_1$ .. $\mathrm{CV}_N\}$, employed to define a reaction coordinate $r(x) = a_0 + \sum_i a_i \mathrm{CV}_i(x)$ as a linear combination with coefficients $\{a_i\}$. The functional form $p = (1+\mathrm{erf}(r))/2$ for the committor probability is also assumed, corresponding to a parabolic approximation of the barrier. The likelihood function is defined as the probability of predicting the observed committment results with the trial RC r:

$$L = \prod_{reac} p(r(x_{reac})) \prod_{prod} [1 - p(r(x_{prod}))] \quad (4)$$

In equation (4), x_{reac} and x_{prod} are initial shooting configurations committed respectively to reactants and products. The likelihood $L(a_0, a_1, ...a_N)$ is maximized to best approximate the true committor function.

Initially we computed the likelihood of single CVs to rank their performance as RC (see the Results section for details). We also optimized combinations of coordination numbers. In the latter case, 12 independent descriptors were considered out of 15 in figure 2, since there is no carbon/carbon coordination and $c(Cl_1,Cl)=c(Cl_2,Cl)$, $c(C,Cl_1)+c(C,Cl_2)=c(C,Cl)$. In the following we use c_1 and c_2 as abbreviations of $c(C,Cl_1)$ and $c(C,Cl_2)$.

When comparing quantitative results from likelihood analysis, we consider as score for each putative RC the ratio between $\log(L)$ and the significant scale $(1/2)\log(N)$, N being the number of observations, called Bayesian information criterion (BIC): when comparing a trial RC including one more CV in the linear combination than another one, the normalized log-likelihood variation must increase more than one unit to be a significant improvement, otherwise the increase is attributed to overfitting.[38,54,56,59]

We set the zero of the log-likelihood axis as a fictitious RC formed by associating random numbers with atomic configurations, hence perfectly decorrelated from the progress of the reaction: we thus define the standard log-likelihood score (SLLS) as

$$SLLS = \frac{\log L(r) - \log L_{\mathrm{random}}}{\frac{1}{2}\log N} \quad (5)$$

We estimated statistical errors due to finite TPS sampling by discarding the first half of accepted TPS shootings (as possibly correlated to the initial shooting point), dividing the second half in two equal parts and analyzing them separately. Applied to each of the data sets A and B, this procedure leads to 4 SLLS scores for each RC tested, and the sampling error is therefore estimated as the standard error (standard deviation divided by $\sqrt{3}$).

Free energy calculations

A series of umbrella sampling[60] simulations were performed, systematically restraining the high-resolution multi-state s path CV with parabolic potentials of the form $\frac{1}{2}k(s - s_0)^2$, starting from initial configurations obtained by either committor analysis or TPS. This corresponds to step VI and VII in figure 1. z is expected to fluctuate around the reference pathway (see e.g. Ref.[42]), however it is important

to detect possible discontinuities along this direction between adjacent US windows, leading to incorrect free energy estimation (see below). An example can be seen in figure 9-(a).

We first adopted 60 US windows equally-spaced by Δs in the s direction and spanning the whole reaction, with $k = k_B T/(\Delta s/2.5)^2$. This choice leads to good overlap between probability distributions in adjacent windows in the case of a flat free-energy profile. In case of local gaps between the distributions, typically due to large variations of the free-energy gradient as found close to the barrier top, the customary procedure consists in adding intermediate windows, possibly with stiffer parabolic potentials (see the Results section for specific cases).

US trajectories are unbiased and combined together into a free energy profile using weighted histogram analysis (WHAM).[61] In order to obtain two independent free-energy estimations with different computer codes, we adopted the codes developed by Alan Grossfield[62] and Andrew L. Ferguson,[63] respectively: given a maximum discrepancy of only 2 kcal/mol along the profiles. We employed 150 bins in s-space and a convergence threshold of 10^{-7} kcal/mol.

The convergence of US sampling was evaluated in the following way. For each window, we discarded the first half as equilibration and used the second half, splitting it again into two equal parts to compute two free energy profiles with WHAM. Statistical error bars on free energy are estimated as the difference between these two profiles. We aimed to reduce the error bars to < 1 kcal/mol, resulting in a duration of about 20 ps for each US window.

Since the free-energy profile is the logarithm of a marginal equilibrium probability density, different free-energy profiles are obtained for a same process when using different CVs.[6] Thermodynamically-meaningful estimates of free-energy differences require therefore the integration of the Boltzmann probability density over the relevant regions in phase space:

$$F_B - F_A = -\beta^{-1} \log \frac{P_B}{P_A} = -\beta^{-1} \log \frac{\int_B ds e^{-\beta F(s)}}{\int_A ds\, e^{-\beta F(s)}} \quad (6)$$

In practice, numerical integration is performed over a regular grid of resolution δs=0.15 in CV space for $s_{12,\mathrm{TPS}}$ and $s_{12,\mathrm{MCA}}$, and $\delta(d_1 - d_2)$=0.2 Åfor $d_1 - d_2$. In the case of the free-energy change of the reaction $\Delta_r F$, regions A and B correspond to the local minima of reactants and products, respectively (the integrals converge quickly with increasing F above each minimum). In the case of barriers $\Delta F^{\ddagger}$, the region of the transition state is defined as the grid bin (of width δs) with the highest F value. The definition is rather insensitive to reasonable variations in the value of δs. Results are reported in Table 2.

Results and Discussion

Exploration of the reaction mechanism

A first exploration of the reaction mechanism is obtained using metadynamics with two-states path CVs s_2 and z_2. The technique excels at quickly escaping from deep local minima, and our choice of CVs – built solely from reactant and product structures – contains no educated guess about the transition state and the reaction mechanism, thus allowing the system to find its own way across the phase space.

The transition happened after about 12 ps (see figure 4), and the simulation was stopped at that moment: in the absence of multiple recrossings producing a stationary probability distribution, the MTD bias is not a quantitative estimate of the free-energy landscape. Nevertheless, the maximum level of the bias accumulated until the transition time provides a first information about the forward barrier of the reaction, typically an overestimation that is less severe the slower the bias deposition. In our case, such maximum level is about 30 kcal/mol, which is indeed not far from the barrier accurately estimated with US, i.e., 23-24 kcal/mol (see below).

Starting from the reactive trajectory obtained with MTD, committor analysis allows to identify transition states among structures sampled in the barrier region. In addition, committor

analysis is based on the generation of unbiased trajectories, which are presumably closer to the MFEP than MTD trajectories. Committor trajectories plotted in figure 4-(b) show the the limitations of s_2 as putative RC. Ideally, all transition state structures should be centered at s_2=1.5, given the symmetric nature of the reaction. Trajectories starting from a MTD configuration at $s_2 \sim 1.5$ on figure 4 (green) indeed have committor probability of 0.46, however, it is possible to find a configuration at $s_2 = 1.6$ with a committor probability of 0.33 (red), whereas a value > 0.5 is ideally expected. This shows that s_2 alone is not a monotonic function of the committor, hence it is a sub-optimal RC.

Shooting points sampled from the MTD trajectory, due to the bias applied in the latter, are not guaranteed to belong to the most likely transition paths that would be sampled by unbiased dynamics. This motivates the TPS step in our protocol: starting from the MTD trajectory, the equilibrium transition state ensemble is systematically targeted, yielding improved information about the reaction mechanism. The results are shown in figure 5: TPS yields a set of transition state configurations that, compared with the MTD candidate, are more symmetrically distributed with respect to the carbon-chlorine coordination numbers c_1 and c_2, as expected.

In principle, different observables might be employed to quantify TPS efficiency, as shown in figure 6. Computing the distance between the initial transition state and the last accepted TPS step in the space of coordination numbers entering the path-CV metric does not allow to rank the different TPS protocols. Conversely, the maximal distance (in the same space) considering all previous TPS steps, as well as the total cumulative time of accepted trajectories, indicate that in AS the optimal interval Δt between configurations is about 10 fs, while SFT appears at least as efficient as the optimal AS and less sensitive to parametrization. We note that the latter criteria correlate with the radius of the transition state ensemble and with the diffusion time therein, respectively.

Observation of plateaus in the distance plots

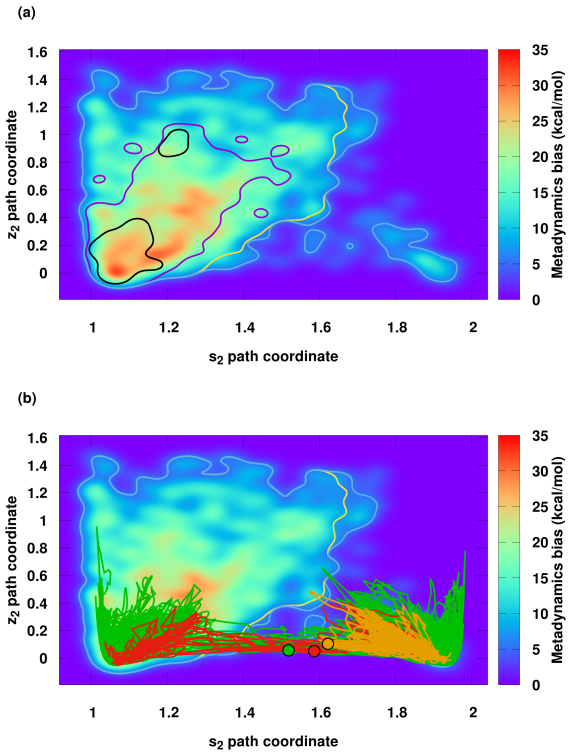


Figure 4: Representation of metadynamics and committor analysis steps of the protocol (IV and V in figure 1), in the (s_2,z_2) plane. The unconverged metadynamics bias at the end of the exploration (12 ps) is represented as a background: In panel **(a)**, contour lines are traced at the level of 5 kcal/mol bias at growing time (1, 4, 11, and 12 ps). In panel **(b)**, only the last two isocontours are represented, together with a sample of committor analysis trajectory. Trajectories originated from green trial point commit to reactants and products with ratio 20:17, indicating a good candidate for transition state, whereas starting from the red or orange point the trajectories are committed to reactants or products, respectively.

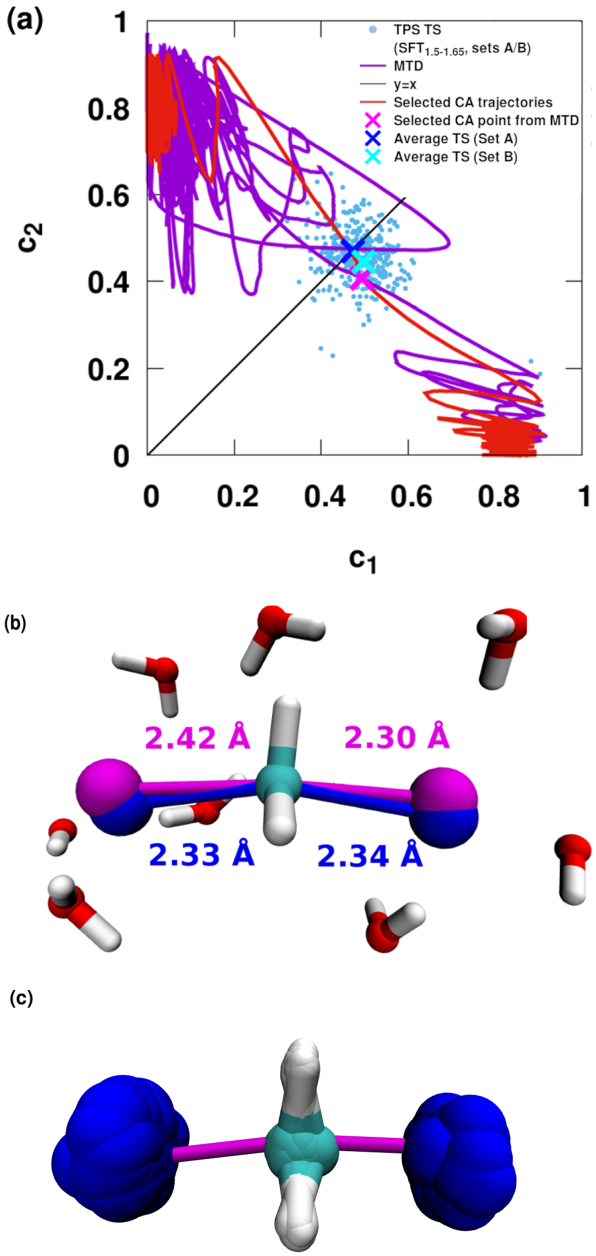


Figure 5: Structural features of transition state configurations obtained from metadynamics and transition path sampling. **(a)** Representation, in the plane of carbon-chlorine coordination numbers, of metadynamics and committor analysis trajectories selected for defining $s/z_{12,MCA}$, accepted shooting points of two sets of transition path sampling simulations ($SFT_{1.35-1.65}$) and corresponding average coordinations. **(b)** Transition state configurations extracted from the committor analysis shooting point (purple) and from the transition state ensemble of data set A in the upper panel (blue). **(c)** Structural variability of the transition state ensembles from data sets A and B.

of figure 6-(a)&(b) suggests that TPS might have reached convergence. To further explore this issue, we performed PCA (based on coordination numbers) and clustering of TPS accepted shooting points: no clear cluster structure could be found, indicating an homogeneous transition state ensemble and thus reaction mechanism (see supplementary information for details). Taken together, these results indicate also that MTD explored transition states close to the optimal ones. PCA of accepted shooting points identify 3 or 4 directions as the most relevant ones: considering maximal plateau values in figure 6-(b) as the radius of the volume spanned by the sampled transition state ensemble, the efficiency ratio between the best and the worst TPS algorithms lies between $(1.0/0.3)^3 \sim 40$ and $(1.0/0.3)^4 \sim 120$.

Reaction coordinates evaluation

We harvested objective information about optimal RCs from the reliable data set formed by TPS trajectories. We remark that intuitive features of the reaction are customarily included in heuristic RC definitions: in particular the mirror-like symmetry between reactants-side and products-side, with the transition state "in the center", suggests the symmetric role of d_1, d_2 or c_1, c_2. However, here we do not assume any knowledge about symmetry and we apply agnostic data-driven approaches to RC identification, to be compared with heuristic RCs (see for instance Ref.[64] for a similar viewpoint in the context of hydrogen-bond definition).

We performed PCA of reactive trajectories sampled with $SFT_{1.35-1.65}$ (data set A), using coordination numbers as input descriptors, to obtain basic insights on the configuration space spanned by the reaction. As shown in figure 7, the first principal component already explains 72% of the total variance. This component, very well correlated with $c_1 - c_2$, could be therefore interpreted as a first approximation of the RC, in agreement with chemical intuition.

The second and third principal components correspond to c(Cl_1,H)-c(Cl_2,H) and c(Cl_1,H) +c(Cl_2,H), and they represent ∼10% each of the total variance. These two components, cap-

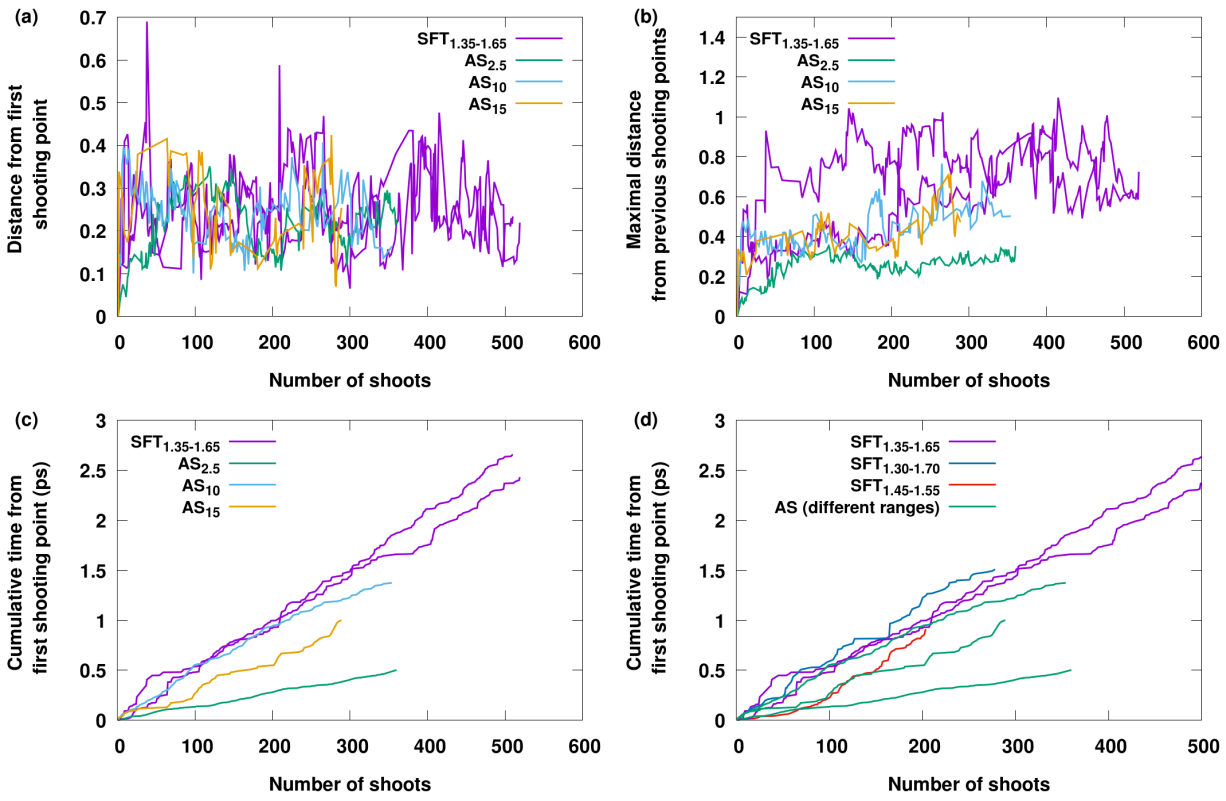


Figure 6: Evolution and convergence of transition states sampled with the AS or SFT algorithms (see Methods). **(a)** Euclidean distance between the first and the last accepted transition states, in the space of coordination numbers, and **(b)** maximal distance between the last accepted transition state and all the preceding ones, as a function of the number of steps. **(c),(d)** Cumulative MD elapsed time between the first and the last accepted transition states as a function of the number of trial shooting points.

turing variations of H-bonds between water and the chloride moiety/ion, indicate that the solvent has varied interactions with the solute and should be studied with care, again in agreement with intuition. However, figure 7-(b) shows also that, due to large fluctuations that overlap together reactants, products and transition states, the solvent-related components might play little role in tracking the progress of the reaction. Indeed, our second component is similar to the "pinching coordinate" proposed as solvent coordinate in ref.,[38] which was found therein to be unable to improve the RC definition.

We also provide a ranking of the set of CV definitions employed in this work, using the SLLS (see equation 5). From figure 8-(a), the best RCs tested are the traditional CV d_1-d_2, and its coordination-equivalent c_1-c_2. Even though there is a considerable scatter between scores obtained from different equivalent parts of the data set, path-CVs s score slightly lower. This result is consistent with PCA and with US results (see below), suggesting that solvent degrees of freedom are not an important part of the RC in the transition path zone. This feature is probably related to the transition state weakly interacting with the polar solvent.

The SLLS for $s_{12,\mathrm{TPS}}$ and $s_{12,\mathrm{MCA}}$ are equivalent (11.4 and 11.1 $\pm$ 1.5) and superior to s_2 (9.1 $\pm$ 1.5), confirming the interest of precisely defined path CVs based on multiple reference structures. As expected, bad coordinates have likelihood close to zero, i.e., close to the score of decorrelated random numbers: this is the case of d_1+d_2, c_1+c_2, and all z path CVs in figure 8-(a). It is interesting to note that CVs based on the a single chlorine atom, such as d_1, d_2, c_1 and c_2, share a similar, very high score of 12.5 $\pm$ 1.5, almost as good as d_1-d_2 and c_1-c_2. This is consistent with PCA results, and denotes a strong correlation between the positions of the two chlorine atoms relative to carbon.

Finally, we optimized the score of linear combinations of coordination numbers employed in the path CVs metric, to perform an agnostic identification of the best RC. SLLS results are shown in figure 8-(b) and a sample of the most relevant optimized collective variables is available in SI.

When a single coordination number is employed, the best RC is c_1, or equivalently c_2 (indicated with oc_1), consistently with PCA as well as intuition. This same result is obtained employing any of the four TPS data sets.

The optimal two-variables combination, oc_2, is formed either as the difference between c_1 and c_2, or between $c(Cl_1, Cl_2)$ and c_1 (the respective scores having no significant difference). In the former case, for instance, the relative weight of c_1 in the linear combination with c_2 is 0.50 ± 0.16, indicating that the intuitive difference $c_1 - c_2$ is indeed recovered by the automatic optimization procedure.

Optimized CVs including more than two coordination numbers do not bring a significant increment in the SLLS score compared to oc_2. This is consistent with PCA results.

The likelihood optimization of the c_1 and c_2 gives the folowing equation for the commitor : $1.209 - 5.526c_1 + 3.737c_2$ which correspond to a value of the commitor of 0.5 for $c_1 = c_2 \approx 0.4$

Free energy profiles

By symmetry, reactants and products must have zero free energy difference: any deviation of computed values from this expected result can be therefore traced back to deficiencies of the statistical sampling, independently from QM approximations. On the other hand, deviations between the computed and the experimental barrier could result both from systematic errors due to QM approximations and from finite sampling.

We performed US with the two different high resolution path CVs as RC, $s_{12,\mathrm{MCA}}$ and $s_{12,\mathrm{TPS}}$, and with the heuristic $d_1 - d_2$ coordinate. Results are presented in figure 9. Integrated free energy values, suitable for experimental comparison, are presented in table 2.

Our data-driven RC built from the most accurate transition path information, $s_{12,\mathrm{TPS}}$, results in an error with respect to experiments of about 2–3 kcal/mol for both the reaction free energy $\Delta_r F$ (expected to be zero) and the barrier $\Delta \mathrm{F}^{\ddagger}$ (expected to be about 27 kcal/mol). After analyzing the distribution of sampled con-

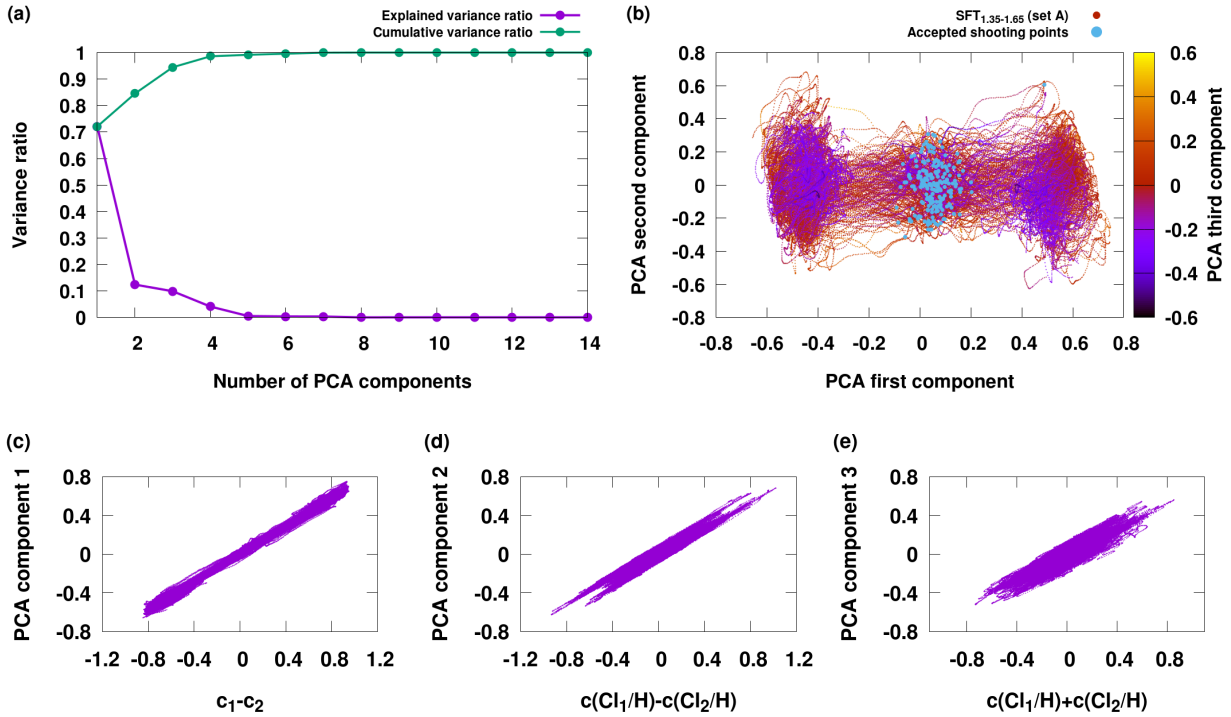


Figure 7: Principal components analysis of coordination CVs in accepted transition path sampling trajectories ($SFT_{1.35-1.65}$ data set A). **(a)** Relative weight and cumulative weight of principal components in terms of variance. **(b)** Accepted transition states and corresponding paths in the space of the three most important components. **(b)**, **(c)** , **(d)** Transition paths projected along each of the three most important components versus heuristic intuitive coordinates.

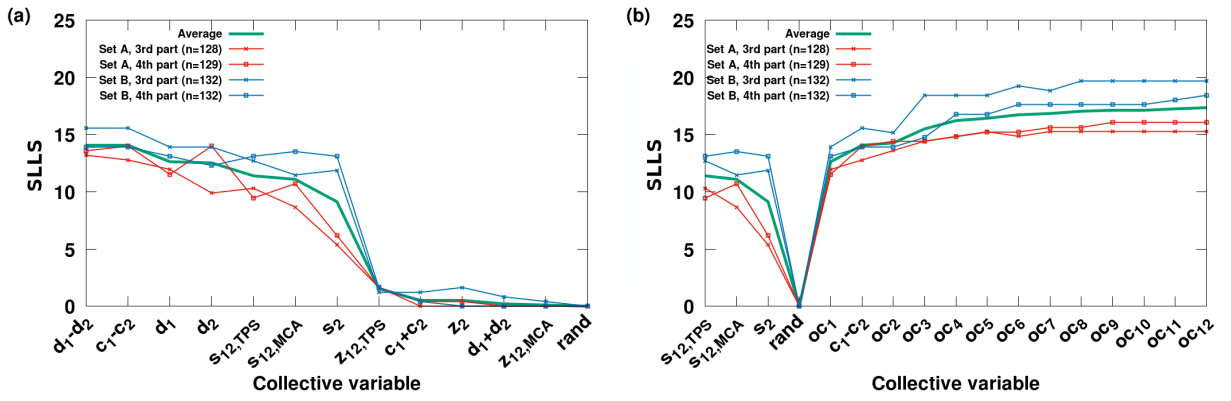


Figure 8: **(a)** Log-likelihood score (SLLS) of 12 CVs, evaluated on transition path sampling shooting points ($SFT_{1.35-1.65}$ data sets A and B). The second half of each data set was cut in two parts, analyzed separately to allow appreciating the statistical convergence of the score.To set the zero level of the score we also include a CV defined as a uniform random variable, hence fully decorrelated from the committor. **(b)** The same score is traced for optimized linear combinations (oc_n) of n coordination number CVs.

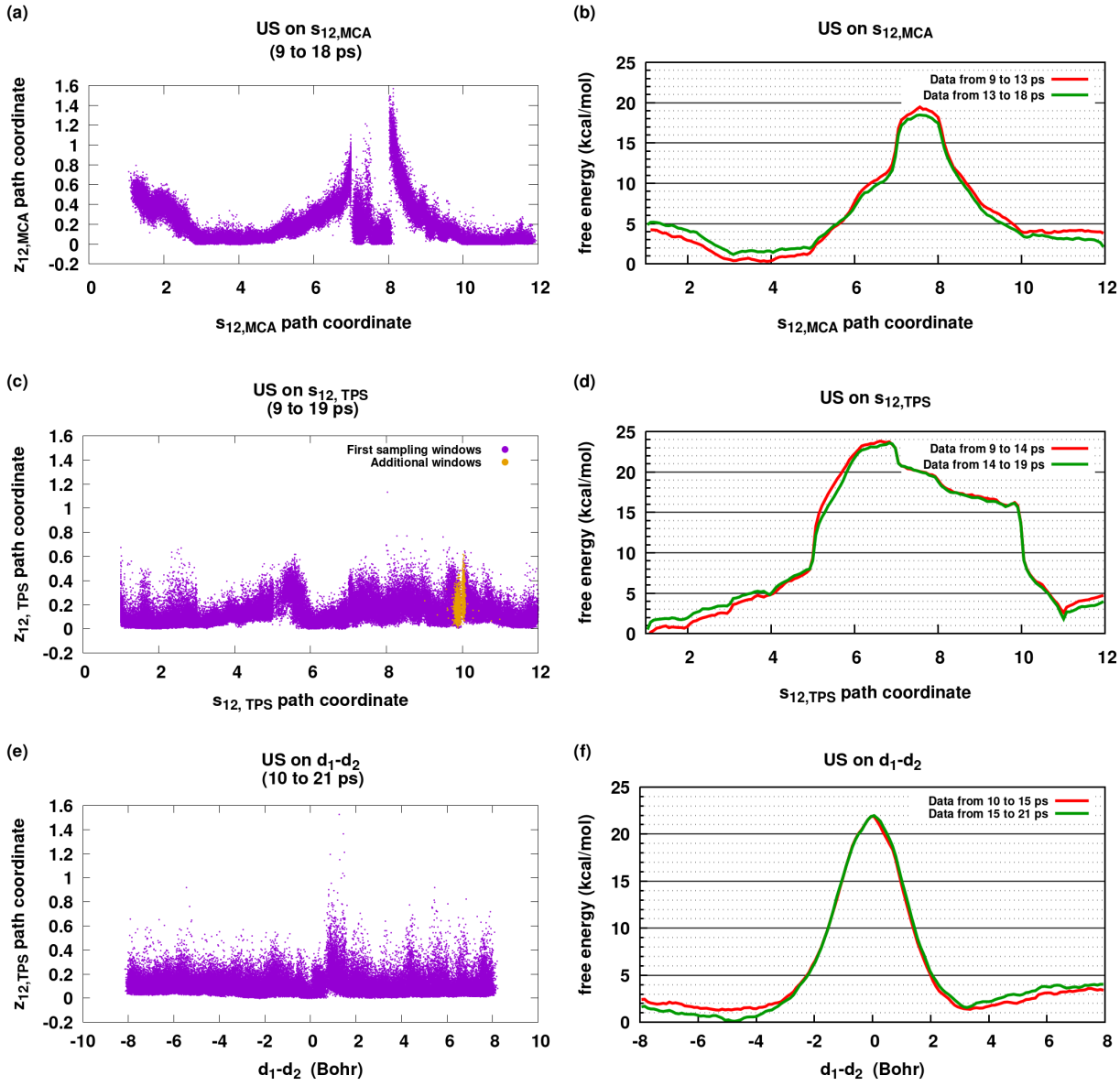


Figure 9: Umbrella sampling results obtained biasing three selected RCs. Left panels **(a),(c),(e)** display sampled points from the second half of simulations (the first half is discarded as equilibration). Right panels **(b),(d),(f)** display free energy profiles computed on the second half of the simulations, divided in two parts to evaluate convergence: discrepancies, indicative of error bars, are within 1 kcal/mol.

figurations in the space of the path-CVS (figure 9-(c)) , as well as in the space of specific coordination numbers (see supporting figures S4 and S5), a small gap has been identified at $s \approx 10$: as customarily done in US, further windows were added in this region, leading to a reliable free-energy calculation. We remark that the stochastic selection of reference structures in the definition of s_{12} can lead to an asymmetric shape of the free-energy profile despite of the symmetric nature of the reaction.

The path CV built from metadynamics-based committor analysis, $s_{12,\mathrm{MCA}}$, results in a sizably underestimated barrier of 19 kcal/mol. Inspection of the s, z space sampled with US displays important gaps, corresponding to large deviations from the putative path (i.e., a significant increase of z). This situation, contrasting with the case of the TPS-based RC that provided a sampling close to its putative path, confirms the intuitive notion about the importance of performing TPS after the metadynamics step and before the US one, in order to build a RC as faithful as possible with respect to spontaneous unbiased transitions. Furthermore, the gaps in US configuration space can also contribute to explain the sizable difference in the barriers estimated with WHAM (that requires only overlap along the biased coordinate, in this case s) for the two RCs.

Compared with $s_{12,\mathrm{TPS}}$, the heuristic RC $d_1 - d_2$ yields an error of only 1 kcal/mol on $\Delta_r F$ but a larger one of about 4 kcal/mol on $\Delta \mathrm{F}^{\ddagger}$. We remark that the error on $\Delta_r F$ is smaller in our work than the 5 kcal/mol value from ref.,[33] where $d_1 - d_2$ was also employed with similar DFT appproximations and 32 water molecules. This could be attributed to more extended sampling in our study: 20 ps (with data analysis after 10 ps of equilibration) for each of 60 umbrella sampling windows, compared to a maximum of 6 ps (with data analysis after 1.5-3 ps of equilibration) for each of 10 values of the reaction coordinate to calculate mean forces in ref.[33]

A recent DFT-MD study[38] reported a free-energy barrier of 15 kcal/mol along $d_1 - d_2$ for the same reaction we studied: the large discrepancy with our own results (as well as with experiments) could be due, at least in part, to the adoption of a QM/MM scheme, while in our opinion it can hardly be attributed to the use of a different exchange-correlation functional (BLYP in Ref.[38] compared with PBE in the present work).

Table 2: Free energy difference between reactants and products $\Delta_r \mathrm{F}$, forward activation free energy $\Delta \mathrm{F}^{\ddagger}$, and maximal free energy difference between points along the profile $\Delta \mathrm{F}_{\mathrm{max}}$. The results of the present work are compared with previous experimental and theoretical studies.

RC used or source	Δ_rF	$\Delta \mathrm{F}^{\ddagger}$	$\Delta \mathrm{F}_{max}$
experiments[12,15]	0	26.5	undefined
theoretical results[33]	5	N.A.	22.2
$\mathbf{d_1}$-$\mathbf{d_2}$	1 ± 0.6	22.6 ± 1	23 ± 0.5
$\mathbf{s}_{12,\mathrm{TPS}}$	2.2 ± 0.9	23.6 ± 0.7	24.3 ± 0.6
$\mathbf{s}_{12,\mathrm{MCA}}$	2.6 ± 1	19 ± 0.8	19.9 ± 1

Inspection of the ensemble of atomic configurations in the plane of s, z path coordinates shows differences in the quality of sampling between $s_{12,\mathrm{TPS}}$ and $s_{12,\mathrm{MCA}}$. We recall that the US bias is applied along the s direction, along which dense sampling is a necessary condition to converge WHAM. On the other hand, the z direction is – by construction – very useful to diagnose the deviation of the sampled pathway from the putative one used in the definition of the RC, as well as to detect potential discontinuities able to render unreliable the reconstructed free-energy profile.

In the case of $s_{12,\mathrm{MCA}}$, US samples a wide range of z values, pointing to a path used in RC definition that is sizably different from the configurations sampled with the help of the parabolic biases. Moreover, figure 9 clearly shows a discontinuity of the US ensemble along the z direction for s slightly above 8. This indicates that the free-energy profile is estimated adjoining two disconnected regions in configuration space, one for each side of the barrier, leading to an unreliable WHAM estimate.

This kind of problems represent one of the main risks of the US technique, and it can only be detected by analyzing the continuity along directions different from the one biased, a task for which the z coordinate is very well suited

as long as the relevant degrees of freedom are included in the metric, eq. 1.

The ensemble sampled with $s_{12,\mathrm{TPS}}$ is instead continuous in the s, z plane (figure 9 c), with small z values all along the pathway, consistently with the expected better quality of the putative path obtained from extensive TPS simulations. However, with our initial set of equally-spaced US windows a small gap in the s-direction (without z disconnection) was still observable near $s_{12,\mathrm{TPS}} = 10$, more clearly displayed by c_1 and c_2 (see Figure5 in SI). As customary done in US, this problem is easily addressed by adding two intermediate windows with stiffer bias on $s_{12,\mathrm{TPS}}$ in the gap region. The additional sampling improves Δ_rF from 3.6 $\pm$ 1 kcal/mol to 2.2 $\pm$ 0.9 kcal/mol, closer to the exact value of zero (table 2). We conclude that the resulting free energy profile is statistically reliable.

Conclusions

In the present work we addressed the challenge of defining reaction coordinates for the computational study of chemical reactions in explicit solvent, using the DFT-level nucleophilic substitution of chloride ion on methyl chloride in water as a test case. At the same time, we propose a practical way out of the following chicken-and-egg paradox: knowledge of the optimal reaction coordinate requires accurate sampling of reactive transitions in phase space, while such sampling needs to be accelerated – to be feasible – by means of techniques that rely on the optimal choice of a reaction coordinate.

Initial reactive pathways are obtained from metadynamics simulations relying on general-purpose path-based collective variables containing information only about reactants and products. Subsequently, transition path sampling is exploited to harvest the most probable transition mechanism in an unbiased fashion. Our results indicate the robustness and cost-effectiveness of this approach, and we advise its systematic use for solution chemistry given the likelihood it gives to sample realistic mechanisms compared to biased sampling of heuristic variables.

At this point, detailed and accurate reaction pathways in phase space are available for extracting an optimal reaction coordinate: on one side, we adopt an algorithm defining a multi-reference path collective variable passing through automatically-selected intermediate configurations. On the other side, we identify relevant degrees of freedom by applying principal component analysis as well as committor-based likelihood maximization techniques in a wide subspace including solute-solute and solute-solvent coordination numbers.

Somehow surprisingly, our results indicate that the solvent does not make an important contribution to the definition of the optimal reaction coordinate, despite making, of course, an important contribution to the energetic of the reaction (strongly increasing the barrier compared to the gas phase). The agnostic optimisation techniques indicate a predominant role of the chlorine-carbon distances, putting on solid quantitative bases a customary hypothesis stemming from chemical intuition.

Careful assessment of different candidate coordinates has been made also via quantitative estimation of free-energy landscapes. To this aim, we exploited extensive umbrella sampling simulations and weighted histogram analysis, reducing statistical error bars to within 1 kcal/mol. We provided evidence that these techniques, despite being widespread, require careful monitoring beyond the biased coordinate to ensure lack of artefacts and proper convergence.

The computed free-energy barriers and differences between reactants and products allow reaching deviations of 2–3 kcal/mol compared to experimentally-derived values when using transition path sampling-based optimal reaction coordinates, not much smaller than using the heuristic $d_1 - d_2$ coordinate.

Even though the present study confirms, to a large extent, the quality of a widespread heuristic coordinate for the specific reaction at hand, we emphasize the importance of defining optimal reaction coordinates in agnostic, data-driven ways for the simulation of generic reactions that lack a large body of published lit-

erature. Our work contributes a flexible toolbox in this direction, that could be applied to a wide spectrum of different reactions without investing extensive work into finding high quality heuristic coordinates

Acknowledgement We gratefully acknowledge Andrea Pérez-Villa and Pauline Bacle for useful discussions and Laura Lupi for assistance with the committor-based likelihood optimization protocol. We gratefully acknowledge the Institute of Computing and Data Sciences (ISCD) from Sorbonne University for funding the PhD thesis of Léon Huet and the MAESTRO project team for insightful discussions and technical support. This work was performed using HPC resources from GENCI (Grants 2020-A0090811069 and 2021-A0110811069).

Supporting Information Available

Supporting Information: Additional calculation details about transition path sampling, clustering, optimized coordinates and umbrella sampling.

References

(1) Pérez de Alba Ortíz, A.; Tiwari, A.; Puthenkalathil, R.; Ensing, B. Advances in enhanced sampling along adaptive paths of collective variables. *The Journal of chemical physics* **2018**, *149*, 072320.

(2) Pietrucci, F. Strategies for the exploration of free energy landscapes: unity in diversity and challenges ahead. *Reviews in Physics* **2017**, *2*, 32–45.

(3) Geissler, P. L.; Dellago, C.; Chandler, D. Kinetic pathways of ion pair dissociation in water. *The Journal of Physical Chemistry B* **1999**, *103*, 3706–3710.

(4) Peters, B. Reaction coordinates and mechanistic hypothesis tests. *Annual review of physical chemistry* **2016**, *67*, 669–690.

(5) Banushkina, P. V.; Krivov, S. V. Optimal reaction coordinates. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2016**, *6*, 748–763.

(6) Jungblut, S.; Dellago, C. Pathways to self-organization: Crystallization via nucleation and growth. *The European Physical Journal E* **2016**, *39*, 1–38.

(7) Orr-Ewing, A. J. Taking the plunge: chemical reaction dynamics in liquids. *Chemical Society Reviews* **2017**, *46*, 7597–7614.

(8) Hughes, E. D.; Ingold, C. K. 55. Mechanism of substitution at a saturated carbon atom. Part IV. A discussion of constitutional and solvent effects on the mechanism, kinetics, velocity, and orientation of substitution. *Journal of the Chemical Society (Resumed)* **1935**, 244–255.

(9) Jonathan Clayden, N.; Warren, S.; Wothers, P. Organic Chemistry by ed. O. 2001.

(10) Bathgate, R.; Moelwyn-Hughes, E. 530. The kinetics of certain ionic exchange reactions of the four methyl halides in aqueous solution. *Journal of the Chemical Society (Resumed)* **1959**, 2642–2648.

(11) Barlow, S. E.; Van Doren, J. M.; Bierbaum, V. M. The gas phase displacement reaction of chloride ion with methyl chloride as a function of kinetic energy. *Journal of the American Chemical Society* **1988**, *110*, 7240–7242.

(12) Heppolette, R.; Robertson, R. The neutral hydrolysis of the methyl halides. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1959**, *252*, 273–285.

(13) Elliott, S.; Rowland, F. S. Nucleophilic substitution rates and solubilities for methyl halides in seawater. *Geophysical research letters* **1993**, *20*, 1043–1046.

(14) Elliott, S.; Rowland, F. S. Methyl halide hydrolysis rates in natural waters. *Journal of atmospheric chemistry* **1995**, *20*, 229–236.

(15) Albery, W. J.; Kreevoy, M. M. *Advances in Physical Organic Chemistry*; Elsevier, 1978; Vol. 16; pp 87–157.

(16) Marcus, R. A. Relation between charge transfer absorption and fluorescence spectra and the inverted region. *The Journal of Physical Chemistry* **1989**, *93*, 3078–3086.

(17) J. McLennan, D. Semi-Empirical Calculation of Rates of 2 Finkelstein Reactions in Solution by a Quasi-Thermodynamic Cycle. *Australian Journal of Chemistry* **1978**, *31*, 1897–1909.

(18) Swain, C. G.; Scott, C. B. Quantitative Correlation of Relative Rates. Comparison of Hydroxide Ion with Other Nucleophilic Reagents toward Alkyl Halides, Esters, Epoxides and Acyl Halides1. *Journal of the American Chemical Society* **1953**, *75*, 141–147.

(19) R. Mathis, J.; Bianco, R.; T. Hynes, J. On the Activation Free Energy of the Cl- + CH3Cl SN2 Reaction in Solution. *Journal of Molecular Liquids* **1994**, *61*, 81–101.

(20) Baesman, S. M.; Miller, L. G. Laboratory Determination of the Carbon Kinetic Isotope Effects (KIEs) for Reactions of Methyl Halides with Various Nucleophiles in Solution. *Journal of Atmospheric Chemistry* **2005**, *52*, 203–219.

(21) Kato, H.; Morokuma, K.; Yonezawa, T.; Fukui, K. A Molecular Orbitals Study of the Nucleophilic Substitution in Methyl Chloride. *Bulletin of the Chemical Society of Japan* **1965**, *38*, 1749–1757.

(22) Gershinsky, G.; Pollak, E. Variational transition state theory for the CI-+ CH3Cl SN 2 exchange reaction in water. *The Journal of chemical physics* **1994**, *101*, 7174–7176.

(23) Grote, R. F.; Hynes, J. T. The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *The Journal of Chemical Physics* **1980**, *73*, 2715–2732.

(24) Bergsma, J. P.; Gertner, B. J.; Wilson, K. R.; Hynes, J. T. Molecular dynamics of a model SN 2 reaction in water. *The Journal of chemical physics* **1987**, *86*, 1356–1376.

(25) Gertner, B. J.; Wilson, K. R.; Hynes, J. T. Nonequilibrium solvation effects on reaction rates for model SN2 reactions in water. *The Journal of Chemical Physics* **1989**, *90*, 3537–3558.

(26) Gertner, B. J.; Whitnell, R. M.; Wilson, K. R.; Hynes, J. T. Activation to the transition state: Reactant and solvent energy flow for a model SN2 reaction in water. *Journal of the American Chemical Society* **1991**, *113*, 74–87.

(27) Ensing, B.; Laio, A.; Parrinello, M.; Klein, M. L. A recipe for the computation of the free energy barrier and the lowest free energy path of concerted reactions. *The journal of physical chemistry B* **2005**, *109*, 6676–6687.

(28) Chandrasekhar, J.; Smith, S. F.; Jorgensen, W. L. Theoretical examination of the SN2 reaction involving chloride ion and methyl chloride in the gas phase and aqueous solution. *Journal of the American Chemical Society* **1985**, *107*, 154–163.

(29) Parker, A. J. Protic-dipolar aprotic solvent effects on rates of bimolecular reactions. *Chemical Reviews* **1969**, *69*, 1–32.

(30) Nibbering, N. M. M. In *Kinetics of Ion-Molecule Reactions*; Ausloos, P., Ed.; NATO Advanced Study Institutes Series; Springer US: Boston, MA, 1979; pp 165–197.

(31) Olmstead, W. N.; Brauman, J. I. Gas-phase nucleophilic displacement reactions.

Journal of the American Chemical Society **1977**, *99*, 4219–4228.

(32) Pellerite, M. J.; Brauman, J. I. Intrinsic barriers in nucleophilic displacements. *Journal of the American Chemical Society* **1980**, *102*, 5993–5999.

(33) Ensing, B.; Meijer, E. J.; Blöchl, P.; Baerends, E. J. Solvation Effects on the SN 2 Reaction between CH3Cl and Cl-in Water. *The Journal of Physical Chemistry A* **2001**, *105*, 3300–3310.

(34) Song, L.; Wu, W.; Hiberty, P. C.; Shaik, S. Identity SN2 reactions X-+ CH3X→ XCH3+ X-(X= F, Cl, Br, and I) in vacuum and in aqueous solution: a valence bond study. *Chemistry–A European Journal* **2006**, *12*, 7458–7466.

(35) Higashi, M.; Truhlar, D. G. Electrostatically embedded multiconfiguration molecular mechanics based on the combined density functional and molecular mechanical method. *Journal of chemical theory and computation* **2008**, *4*, 790–803.

(36) Zheng, H.; Wang, S.; Zhang, Y. Increasing the time step with mass scaling in Born-Oppenheimer ab initio QM/MM molecular dynamics simulations. *Journal of computational chemistry* **2009**, *30*, 2706–2711.

(37) Tirado-Rives, J.; Jorgensen, W. L. QM/MM Calculations for the Cl–+ CH3Cl SN2 Reaction in Water Using CM5 Charges and Density Functional Theory. *The Journal of Physical Chemistry A* **2019**, *123*, 5713–5717.

(38) Leitold, C.; Mundy, C. J.; Baer, M. D.; Schenter, G. K.; Peters, B. Solvent reaction coordinate for an SN2 reaction. *The Journal of Chemical Physics* **2020**, *153*, 024103.

(39) Bailleul, S.; Dedecker, K.; Cnudde, P.; Vanduyfhuys, L.; Waroquier, M.; Van Speybroeck, V. Ab initio enhanced sampling kinetic study on MTO ethene methylation reaction. *Journal of Catalysis* **2020**, *388*, 38–51.

(40) de Alba Ortíz, A. P.; Vreede, J.; Ensing, B. *Biomolecular Simulations*; Springer, 2019; pp 255–290.

(41) Pietrucci, F.; Saitta, A. M. Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios. *Proceedings of the National Academy of Sciences* **2015**, *112*, 15030–15035.

(42) Magrino, T.; Pietrucci, F.; Saitta, A. M. Step by Step Strecker Amino Acid Synthesis from ab Initio Prebiotic Chemistry. *The Journal of Physical Chemistry Letters* **2021**, *12*, 2630–2637.

(43) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **1996**, *77*, 3865.

(44) Hutter, J.; Curioni, A. Car–Parrinello Molecular Dynamics on Massively Parallel Computers. *ChemPhysChem* **2005**, *6*, 1788–1793, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cph

(45) Troullier, N.; Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Physical review B* **1991**, *43*, 1993.

(46) The PLUMED consortium, Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* **2019**, *16*, 670–673.

(47) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A., *et al.* PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009**, *180*, 1961–1972.

(48) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy

space. *The Journal of chemical physics* **2007**, *126*, 054103.

(49) Pérez-Villa, A.; Pietrucci, F.; Saitta, A. M. Prebiotic chemistry and origins of life research with atomistic computer simulations. *Physics of life reviews* **2018**,

(50) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(51) Bussi, G.; Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics* **2020**, *2*, 200–212.

(52) Dellago, C.; Bolhuis, P. G. Transition path sampling and other advanced simulation techniques for rare events. *Advanced computer simulation approaches for soft matter sciences III* **2009**, 167–233.

(53) Bolhuis, P. G.; Swenson, D. W. Transition Path Sampling as Markov Chain Monte Carlo of Trajectories: Recent Algorithms, Software, Applications, and Future Outlook. *Advanced Theory and Simulations* **2021**, *4*, 2000237.

(54) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *The Journal of chemical physics* **2006**, *125*, 054108.

(55) Peters, B.; Beckham, G. T.; Trout, B. L. Extensions to the likelihood maximization approach for finding reaction coordinates. *The Journal of chemical physics* **2007**, *127*, 034109.

(56) Mullen, R. G.; Shea, J.-E.; Peters, B. Easy transition path sampling methods: Flexible-length aimless shooting and permutation shooting. *Journal of chemical theory and computation* **2015**, *11*, 2421–2428.

(57) Jung, H.; Okazaki, K.-i.; Hummer, G. Transition path sampling of rare events by shooting from the top. *The Journal of chemical physics* **2017**, *147*, 152716.

(58) Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(59) Kass, R. E.; Raftery, A. E. Bayes factors. *Journal of the american statistical association* **1995**, *90*, 773–795.

(60) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.

(61) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of computational chemistry* **1992**, *13*, 1011–1021.

(62) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.11. `http://membrane.urmc.rochester.edu/?page_id=126`, Last accessed 28 September 2022.

(63) Ferguson, A. L. BayesWHAM: A Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. *Journal of Computational Chemistry* **2017**, *38*, 1583–1605, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.

(64) Gasparotto, P.; Ceriotti, M. Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond. *The Journal of chemical physics* **2014**, *141*, 174110.

4.2 Conclusion in the context of the Thesis

This work was the concluding piece of Théo Magrino's thesis. For me, it served as an introduction project to the tools I would use in future research. In the paper, we established a comprehensive agnostic protocol that allows us to generate a reaction coordinate by following a workflow where the only initial inputs are the definitions of the reactants and products of the reaction.

For the remainder of this thesis, three major outcomes must be emphasized:

- The use of the Transition Path Sampling dataset to generate a PCV is more efficient than using data from a committor analysis alone. The resulting PCV exhibits a more symmetric behavior with respect to the studied reaction and the free energy curve. The final barrier height is also closer to the experimental value and the results with the heuristic RC.

- The Standard Log-Likelihood Score (SLLS) appears to be a good indicator of the quality of the CV at the top of the barrier.

- Despite our efforts to generate a reliable PCV the heuristic variable Δ d behaves almost perfectly during the umbrella sampling and presents similar results to the SLLS. Thus, it remains a better reaction coordinate than our database-derived one.

The first statement will influence the next article that I present. In the next article, a large part of the protocol has been enhanced based on the observations we made in this critical assessment.

The second and third outcomes have implications for the last article presented in this work, in which we decided to study the possible inference of kinetics on the same chemical system. For that, we decided to retain SLLS as a promising quality criterion and center our study on the Δd variable, which, despite being a heuristic variable, remains the best available option for our study.

Chapter 5

Application to a prebiotic system

5.1 Presentation of the article

During his Ph.D., Theo Magrino conducted a study on the commonly accepted glycine prebiotic aqueous synthesis: the Strecker mechanism[14]. In parallel, he also performed metadynamics explorations to sample the full pathway without any prior assumptions about the intermediates. Among all of his attempts, only one has successfully completed a full synthesis from the Strecker reactants to glycine in a single trajectory. Interestingly, this pathway differed from the Strecker mechanism and presented notable features for prebiotic chemistry. This new pathway is the basis for this article. We conducted a study to check the observed intermediates, validate the transition states, and calculate the free-energy curve associated with each step of this new synthesis.

However, an issue arose when the final measured free energy difference between the Strecker reactants and glycine appeared to be inconsistent between the two studies. The discrepancy was of 10 kcal.mol^{-1}, which represented one fourth of the total free energy difference he determined for glycine. The origin of this discrepancy comes from the original study of the Strecker pathway, as it appeared that the data contained some hysteresis. Correction of these hysteresis has been submitted in an erratum and is currently under publication process (see Appendix D).

In the article on the oxyglycolate pathway, the name we gave to the new pathway, we decided to perform the quantification step using a NNP. This was carried out as a collaborative effort between a senior Ph.D. student, Timothée Devergne, and me. He specialized in training and sampling the NNP and also performed a part of the AIMD calculations. We agreed that the implementation of the NNP should be conducted completely independently of the AIMD simulation to increase our exigences from his previous studies[6, 57]. We executed the MLMD protocol for each US step. To verify the presence of errors in the machine learning results, we recalculated three of the five steps using a fully AIMD framework.

The article presented here is not in its final version, as it has recently been accepted with minor revisions by the Journal of Physical Chemistry Letters. The modifications

suggested by the referees should be addressed soon. The supporting information for the article is available in the Appendix C.

A new route to the prebiotic synthesis of glycine via quantum-based machine learning calculations

Léon Huet,[†] Timothée Devergne,[†,‡,¶] Théo Magrino,[†] and A. Marco Saitta[*,†]

†*Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS UMR 7590, Paris 75005, France*

‡*Atomistic Simulations, Italian Institute of Technology, 16142 Genoa, Italy*

¶*Computational Statistics and Machine Learning, Italian Institute of Technology, 16142 Genoa, Italy*

E-mail: marco.saitta@sorbonne-universite.fr

Abstract

In this work, we study the synthesis of glycine, the simplest amino acid, using *ab initio* molecular dynamics and enhanced sampling techniques to explore and quantify novel potential pathways. Our protocol integrates state-of-the-art machine learning approaches, allowing us to sample relevant chemical spaces more efficiently. We discover a novel 'oxyglycolate path,' distinct from the "standard" Strecker mechanism, identify new intermediates, and provide a full thermodynamic characterization of all reaction steps. This alternative pathway aligns better with meteoritic and experimental observations, paving the way for further investigations. Integrating quantum accuracy and machine learning in prebiotic chemistry represents a methodological milestone advancing the exploration of life's prebiotic origins.

In the field of prebiotic chemistry, the synthesis of glycine holds a crucial role. Being the simplest amino acid and a primary component of proteins, understanding its pre-life synthesis might shed light on the formation pathways of other amino acids. Several studies reveal

glycine's presence in meteorites,[1–3] yet its existence in the interstellar medium (ISM) remains unresolved.[4] A leading hypothesis suggests that inorganic materials and simple molecules, whether in icy chondrites, meteorite cores, or primordial Earth's oceans, might have created favorable environments for glycine synthesis.[5] This idea is central to the Strecker mechanism,[5] which involves "simple", fundamental molecules[6]—water, formaldehyde, ammonia, and hydrogen cyanide— detected in both ISM and meteorites.[7,8] It relies on the formation of glycinonitrile (aminoacetonitrile) as first main intermediate by the successive addition of amine and cyanide on the formaldehyde molecule, then the nitrile bond is replaced by a carboxylic group in a two step hydrolysis. Glycinonitrile and glycinamide are generally considered as the signature intermediates of this pathway. S.L. Miller postulated and tested this mechanism[5] based on his landmark 1950s experiments with Urey.[9] These results kick-started research into amino acids' origins on Earth and beyond. However, the discovery of β- and γ-amino acids in meteorites[10] introduced alternative synthesis pathways to reconcile observations the Strecker mechanism alone could not address.[11]

In challenging observational and experimental contexts, *ab initio* computational modeling based on quantum mechanics becomes indispensable. It provides rigorous insights into amino acid synthesis mechanisms and complements experimental data, especially in prebiotic chemistry/origins of life (PCOL) studies where observational conditions are often elusive. Theoretical exploration of synthesis pathways helps understand the synthesis conditions of critical compounds and discover new pathways, enhancing our understanding of early Earth prebiotic chemistry. Advanced theoretical tools have been developed to ensure reliability and transferability in such studies.[12] Increased computational power has made complex simulations more accessible, enabling accurate electronic structure estimations using Density Functional Theory (DFT). Following a detailed examination of the Strecker synthesis,[13] we applied and augmented our computational tools to the Strecker reactants, seeking alternative pathways. This work led to the identification of a novel synthesis route, designated as the 'oxyglycolate path' in this manuscript.

In our previous paper,[13] the estimation of potential and free-energy was entirely driven by DFT, through *ab initio* Molecular Dynamics (AIMD). Our protocol[14] consists of three main steps:

1. An exploration step that enables the discovery of reactive pathways;
2. A verification step of intermediates and putative transition states;
3. A quantification step that ensures the accurate estimation of the free energy landscape for all the elementary steps of the reaction. This last step represent the most expensive part of our work in CPU hours.

This protocol, essential for ensuring thermodynamic and statistical accuracy, is extremely costly and lengthy, from the computational point of view, for reactions occurring in the condensed phase. It has thus far *de facto* impeded significant progress in PCOL ab initio studies, particularly in the crucial case in which reactions occur at the water/mineral interfaces.

In the present work in particular we have devised a novel methodology to address this inherent limitations of AIMD. One major constraint is the computational scalability of these simulations, limiting their application to systems predominantly at the nanosecond/nanometer scale. In response, we have developed an optimized procedure that combines the accuracy of AIMD with state-of-the-art in-house enhanced sampling techniques and recent machine learning approaches. This clever combination allows us to extensively map reaction free energy landscapes during the third step of our protocol, while keeping the same accuracy as AIMD, as we demonstrated in Ref([15]).

We will refer in the following to this as Machine Learned Molecular Dynamics (MLMD), positioning it as a complement to "full/traditional" AIMD. All the technical details are presented in the "Methods" section at the manuscript conclusion. The effectiveness and scalability of this tool make it, we believe, indispensable for future research in prebiotic chemistry. To the best of our knowledge, this study represents one of the first applications,[16] and the first fully agnostic one in terms of CV choice, that combines fully validated and

fully autonomous uses of MLMD in PCOL studies. This sets a pioneering standard for the future development of reliable machine learning potentials in condensed phase/mineral surface chemistry.

This manuscript thus presents an extensive study integrating two significant advancements: the identification of a previously unreported glycine synthesis pathway, and a first application of our augmented protocol within the context of PCOL studies. This research builds upon, and advances beyond, the methodologies established by our previous works.[13,15]

In the following, we provide a detailed presentation of those results, including our analysis of the thermodynamic and chemical implications derived from our findings.

The final mechanism from the explorative step is illustrated in Figure 1. Initially, hydrogen cyanide **R** is activated via an acid-base reaction. Next, the cyanide anion (**1**) adds to formaldehyde, forming cyanomethanolate (**2**), which reacts with the solvent to produce glycolonitrile (**2'**). Glycolonitrile undergoes a two-step hydration-hydrolysis, forming the unstable oxiranimine (**3**) and then hydroxyacetamide (**4**). An addition-elimination step yields glycolic acid (**4'**), which reacts with the solvent to become glycolate (**5**). Finally, an S_N2 reaction at the carboxylate group's α position produces the glycine anion (**P**). Figure 2 shows the full trajectory from the initial "Strecker" precursors to glycine.

Among all the intermediates in the mechanism, two were discovered after the initial exploration and were subsequently incorporated into the set during the validation step: **2'** and **4'**.

- Cyanomethanolate **2** underwent spontaneous reaction with the solvent after 1.5 ps of unbiased trajectories, leading to the inclusion of **2'**, as parasitic intermediate, in the initial set.

- While sampling the transition state of the initial (**4** $\rightarrow$ **5**) step, the trajectories became trapped in a new intermediate state, **4'**, which is the conjugate acid of **5**.

The addition of these two intermediates to the initial set highlights the importance of

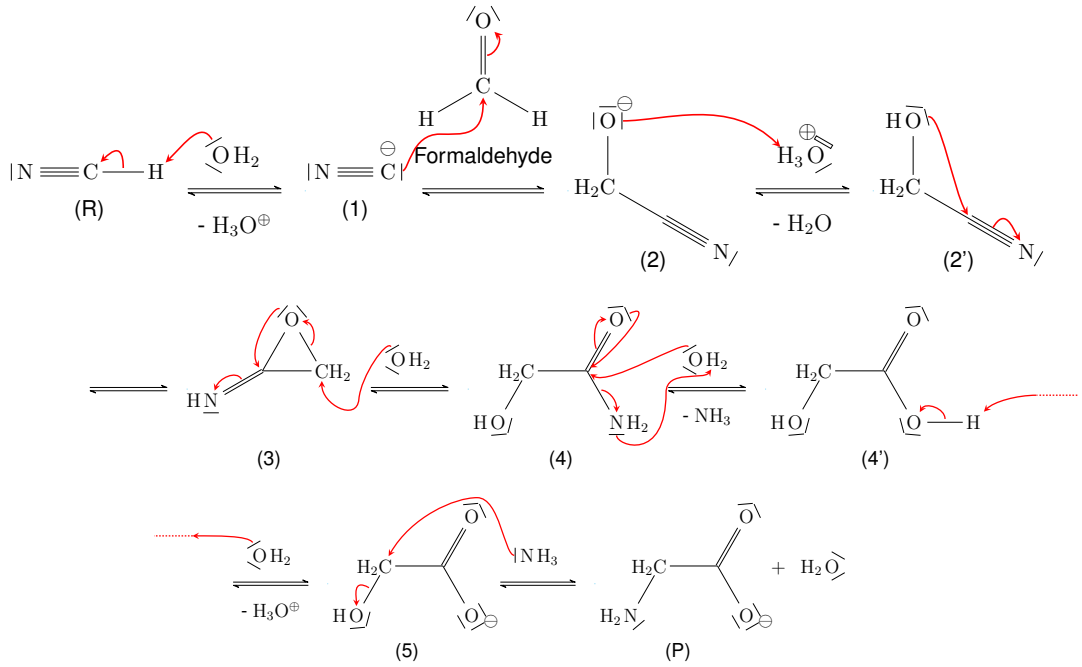


Figure 1: Mechanism of the oxy-glycolate path, decomposed in elementary acts observed during the exploration step (metadynamics and transition path sampling). All electron displacements are schematically represented by red arrows. Three proton transfers, between (2') to (3), (3) to (4) and (5) to (P), are not represented for readability but are still followed by our protocol as part of the elementary acts.

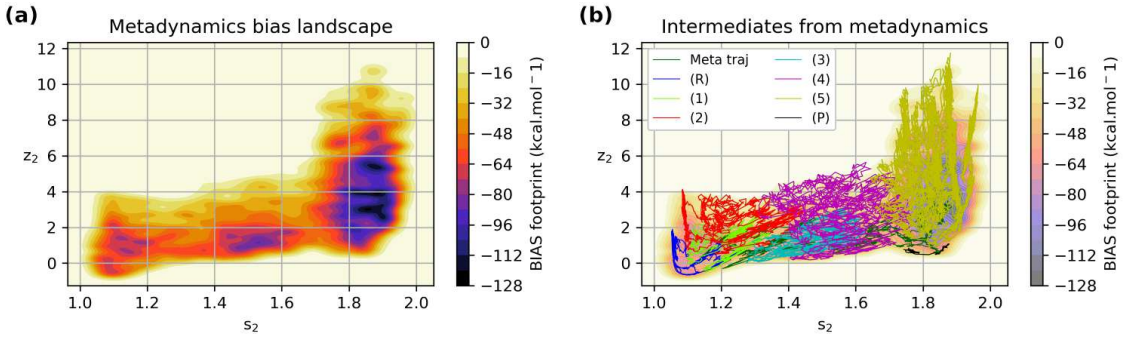


Figure 2: 2D graphics of the metadynamics trajectory that breaks toward glycine. **(a)** The footprint of the bias introduced in metadynamics. **(b)** The full trajectory colored with the initial set of intermediates out of the molecules observed during the metadynamics, this initial set is checked and validated by further investigation to avoid human and sample errors.

caution in conducting biased explorative trajectories. Our method's reliability is supported by the meticulous consideration and verification of the quality of our initial landscape exploration during the subsequent steps of the protocol.

The primary steps of the mechanism were quantified with three different case:

i Three simple proton transfers with the solvent were determined only with AIMD.

ii All other complex steps were elucidate with MLMD.

iii Additionally, the curves for the steps **2'** $\rightarrow$ **3**, **3** $\rightarrow$ **4**, and **5** $\rightarrow$ **P** were independently calculated using AIMD to assess the reliability of MLMD.

The resulting curves from these calculations are presented in Figure 3. Except for oxiranimine, which exhibits the highest intermediate free energy, the differences between AIMD and MLMD curves consistently remain within 0.5 kcal·mol^{-1} for each intermediate and within 3 kcal·mol^{-1} for the transition states. Moreover, the shapes of the curves are well-reproduced.

A comprehensive overview of the relative free energy balance is reported in Figure 4. The pathways using mostly AIMD and mostly MLMD align closely with an overall balance difference of 0.1 kcal·mol^{-1}, resulting in a value of -21.7 kcal·mol^{-1} for the free energy of the glycine anion. This pathway represents, to our knowledge, the first fully *ab initio* discovered route for glycine synthesis, combining *ab initio* DFT and machine learning potentials.

The remarkable agreement between AIMD and MLMD outcomes attests to the efficiency of the neural network potential. AIMD and MLMD quantitative calculations were conducted independently, sharing only the same variables and initial geometries—a noteworthy achievement.

Available experimental data are compared to our values in Table 1. All simulations achieved a convergence error below 0.5 kcal·mol^{-1} per step. For the first five steps, our data shows less than a 10 kcal·mol^{-1} difference from experimental values, within the DFT method's standard error ($\approx$10 kcal·mol^{-1},[25] see Methods). For the last three steps, larger differences may stem from experimental biases, PBE systematic errors, and accumulated

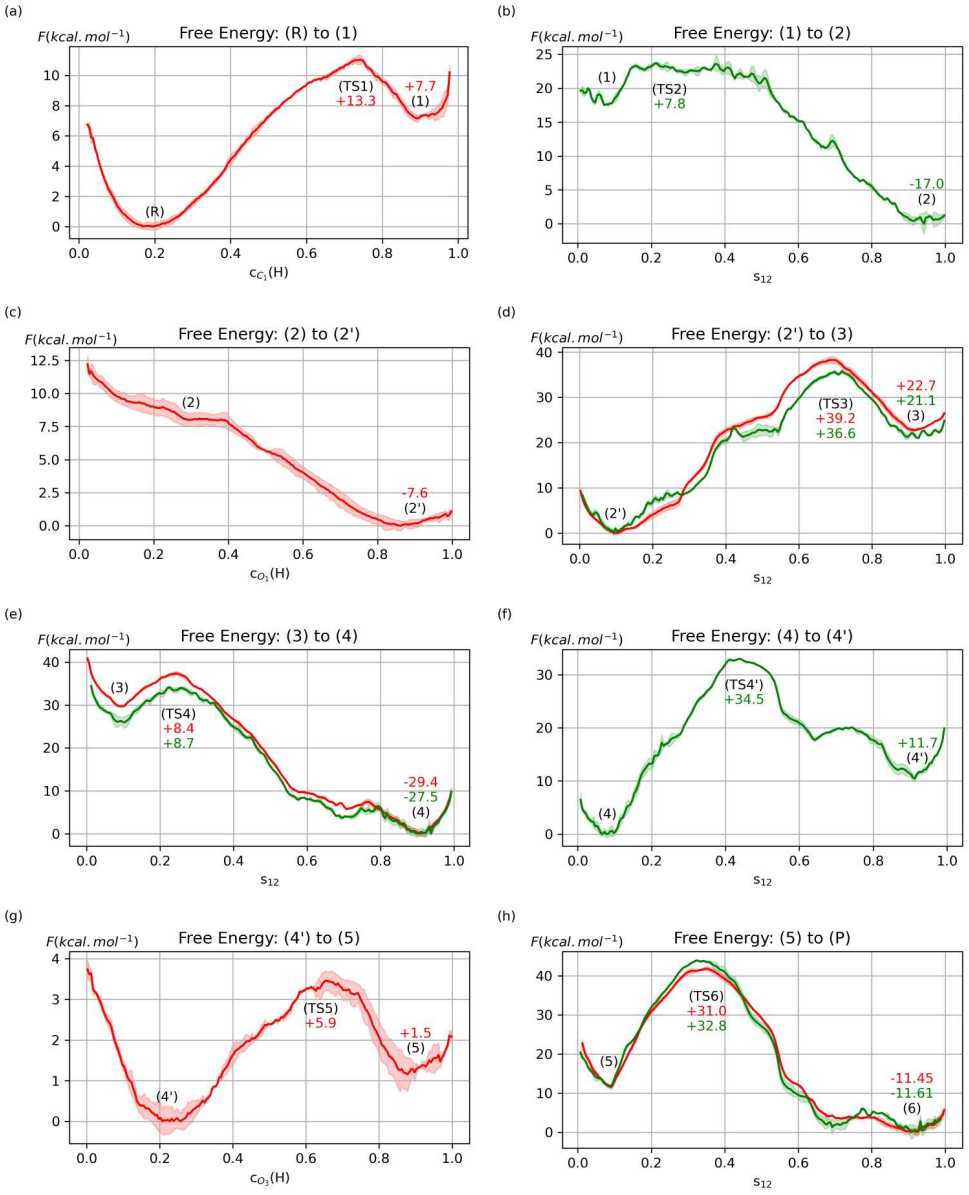


Figure 3: Free energy profiles for our complete mechanism, with AIMD results in red and MLMD results in green. Energetic gaps were calculated by integrating the free energy over the minima of reactants and products. Intermediates are numbered according to the mechanism in Figure 1. The s_{12} variables are data-driven from transition path sampling data (see methods section). The $C_X(H)$ variables represent hydrogen coordination for relevant atoms during protonation/deprotonation steps. All reaction coordinates, including coordination variables, were linearly re-scaled from 0 (reactants) to 1 (products) for consistency.

Table 1: Comparison of the Gibbs free energy from the literature with the Helmholtz free energy obtained from our NVT calculations. (in kcal.mol^{-1})

Symbol	Species	G_{aq} exp.	REF	F_{aq} theo.	REF	Δ
R	Strecker reactants	0 (ref. level)		0		0
1	cyanide ion	+12.7	[17]*	+7.7		-5.0
2	cyanomethanolate	NA		-9.3		NA
2'	glycolonitrile	-12.2	[18]	-16.8		-4.7
3	2-oxiranimine	(+15.0 only ΔH)	[19]**	+4.2		-10.8
4	hydroxyacetamide	-12.9	[20,21]***	-23.2		-10.3
4'	glycolic acid	NA		-11.6		NA
5	glycolate	NA; (+5.2 wrt **4'**)	[17]*	-10.0		NA
P	glycine anion	-7.7 (+13.3 wrt **P'**)	[22]	-21.7		-14.0
P'	Z-glycine	-21	[23]	-35.0 (-13.3 wrt **P**)	[22]	-14.0

* ΔG value obtained from the pKa of the acid-base pair.

** Theoretical estimation.

*** This experimental estimation carries an unspecified uncertainty, it is still presented here as the sole indication of an experimental result to the best of our knowledge. Other theoretical data, in line with our glycolonitrile value, suggest that hydroxyacetamide should be at -37.1 kcal·mol^{-1}.[24] This value remains significantly distant from the experimental result.

The colors represent the last step type: red for AIMD and green for MLMD. We compare experimental Gibbs free energy with theoretical Helmholtz free energy, leveraging the minimal volume variation in the isothermal–isobaric ensemble for chemical reactions in the condensed phase.

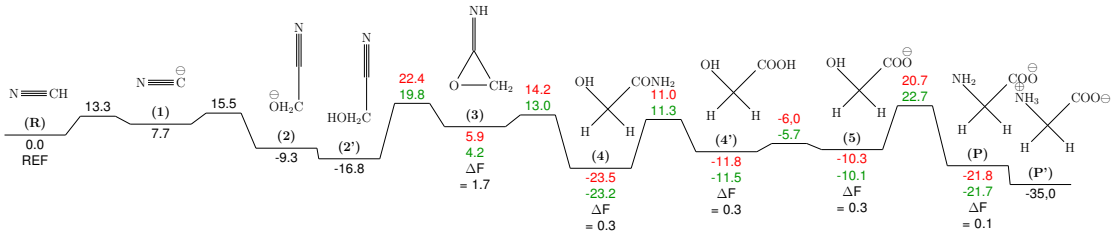


Figure 4: Balance of the Helmholtz free energy for all steps of our "non-Strecker" mechanism. Green values refer to MLMD results for complex steps (with s_{12} as the reaction coordinate) and AIMD for proton transfers. These values are used throughout the article. Red values are obtained when prioritizing AIMD, especially where both AIMD and MLMD have been conducted. This presentation illustrates the concordance between the two methods in the overall free energy balance of the mechanism.

convergence errors. Our results closely align with other theoretical studies on these compounds.[24,26]

In the following we provide a critical discussion of our proposed mechanism in comparison to the Strecker pathway using our previous article as reference.[13] This reference initially exhibited some discrepancies, which have since been rectified through a correction.[27] We adopt the revised general diagram as our reference. The final free energy difference in the correction aligns very well with the results presented here.

To the very present state The Strecker pathway continues to exhibit lower intermediates in free energy and lower activation barriers compared to the pathway we are proposing. In quantitative terms, this translates to -42.1 kcal·mol^{-1} for the glycine-amide intermediate in the Strecker pathway, in contrast to -23.2 kcal·mol^{-1} for the glycol-amide intermediate and 31 kcal·mol^{-1} for the last and highest barrier in the Strecker pathway, as opposed to 36.6 kcal·mol^{-1} for the oxiranimine formation barrier in our proposed mechanism.

These two statements suggest that under standard conditions, the Strecker mechanism is more likely to occur. However, despite numerous attempts, the Strecker pathway does not manifest during the metadynamics step of our process (refer to the method section), despite its apparent agnosticism. To our knowledge, this could be attributed to two non-mutually exclusive factors:

- Our s_2 and z_2 may struggle to resolve details of a multi-step pathway in a high-dimensional free-energy landscape, which includes at least two glycine synthesis pathways. Enhancements in the metadynamics step, particularly on variable selection and solvent degrees of freedom, are needed while maintaining the agnostic nature of our variables.

- The Strecker pathway may be kinetically blocked in our metadynamics simulation due to long equilibration times involving solvent degrees of freedom, which are challenging to accelerate with an external bias. This suggests the need for further investigation to determine if this kinetic blockage can be overcome with alternative variables. If so, this blockage would likely not imply kinetic effect on the Strecker pathway experimentally.

This statement, along with the associated uncertainty, must be considered in light of the potential advantages that this new pathway could offer in advancing our understanding of the prebiotic synthesis of glycine.

The primary distinction from the Strecker mechanism is that this new pathway can be activated by a basic catalyst. Steps $(R) \rightarrow (1)$, $(3) \rightarrow (4)$, $(4) \rightarrow (4')$, and $(4') \rightarrow (5)$ are more favorable in a basic environment, involving acid-base equilibria toward basic compounds or substitution reactions facilitated by HO^-. Additionally, step $(2) \rightarrow (2')$ can be seen as a parasitic reaction less favorable in a basic environment, potentially lowering the barrier for intermediate (3) to around 30 kcal·mol^{-1}, comparable to the limiting reaction of the Strecker pathway.[13] Given the supposed primitive ocean's pH in the [8-9] range,[28,29] this new pathway is more consistent with early Earth conditions.

The first major intermediate in this new pathway is glycolonitrile. In the literature, the formation of glycolonitrile is typically considered only as a parasitic reaction in the synthesis of glycine.[28] Notably, the formation of glycolonitrile is frequently seen as competing with the formation of glycinonitrile (aminoacetonitrile), one of the principal intermediates in the Strecker synthesis. Our calculations indicate that the pathway to glycolonitrile is highly favorable ($\Delta F = -16.8 \text{kcal} \cdot \text{mol}^{-1}$), though slightly less so than glycinonitrile (-22.3

kcal $\cdot$ mol^{-1} [13]). Glycolonitrile requires fewer steps, has lower barriers, and is more stable than the intermediates towards glycinonitrile, suggesting it is a favorable kinetic product in water. An accelerated hydrolysis step thanks to environmental factors could make the "oxy-glycolate" path even more competitive with the Strecker one.

In our calculations, the hydration-hydrolysis of glycolonitrile begins with the formation of 2-oxiranimine, an exceedingly unstable intermediate, suggesting the possibility of more efficient pathways. However, In Jammot's experimental papers,[20,21] the reaction is catalyzed by a borate buffer, likely creating a cycle involving glycolonitrile's carbons in an intramolecular activation step, similar to our proposed pathway. Despite uncertainties in the oxyglycolate path, the free energy balance between glycolonitrile and hydroxyacetamide remains thermodynamically correct. While oxiranimine could react with a solvated amine to form aminoacetamide (a Strecker pathway intermediate), this would require a high amine concentration, making the production of hydroxyacetamide and glycolic acid more probable.

Glycolic acid has been detected in meteorites[30,31] and is of ongoing theoretical and experimental interest.[32,33] Hydroxyacetamide, synthesized in prebiotic experiments by UV irradiation of interstellar ice analogs alongside glycolic acid,[34] has been theoretically proposed to have a synthesis pathway similar to ours,[35] with a half-life of 52 years at 25°C. Their presence in this new pathway toward glycine could explain their relative abundance with glycine in meteorites,[1–3] making the thermodynamic data of this study valuable for further investigations. Additionally, glycolic acid is a known product of glycine's hydrothermal decomposition.[36–38]

The transition to glycolate before progressing towards glycine could be contradictory with respect to chemical intuition. To gain further insights into this important step, we chose to specifically investigate whether this final step could be more favorable without the glycolate intermediate. Our findings indicate that, in our simulations, the basic form yields to a more stable transition state compared to the acidic form (See in supporting information **S-4**).

Studies on the decomposition of glycolonitrile suggest that this molecule could have been synthesized in the ISM[39] and incorporated into chondrites, where the remainder of the mechanism might have taken place. This implies an environment much more complex than the one we simulate, calling for further investigations.

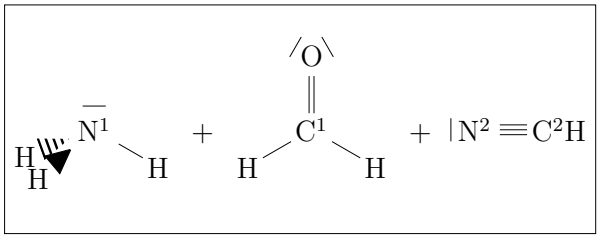
Conclusions

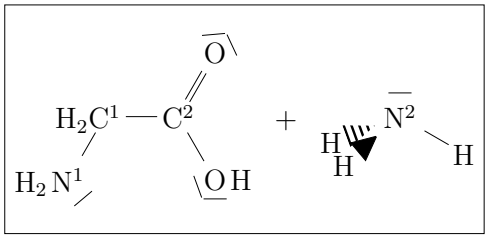
In this work, we conducted the first fully-agnostic one-pot *in silico* study of the aqueous synthesis of glycine, the simplest amino acid. Using the same precursors as the Strecker reaction, our accelerated exploration revealed a novel pathway, discovering new intermediates and providing a comprehensive thermodynamic characterization. This new pathway appears as a realistic alternative to the Strecker mechanism, in agreement with meteoritic sample observations. Our study brings insightful data on the thermodynamics of its intermediates.

We used a machine-learning-driven potential to accurately calculate free-energy profiles, reducing the computational cost by nearly an order of magnitude. This study, combining *ab initio* accuracy with advanced sampling and machine learning in PCOL, aims to open new research perspectives, scaling up system sizes and timescales to better match experimental conditions and model complex prebiotic synthesis processes.

Methods

Our computational protocol needs only as input the definition of reactants and products of the full mechanism. It combines state of the art enhanced sampling techniques and data based Collective Variables (CV). All AIMD calculations have been conducted with CPMD.4.3 software.[40] The enhanced sampling methods were performed using a modified version of the Plumed1.3 plugin of CPMD,[41] the modified source code is freely available on demand, or online at this page: https://sites.google.com/site/fabiopietrucci/. The protocol is composed of three major step:





	Elements σ			
	C	O	N	H
C^1	0.0	1.2	0.0	2.0
C^2	0.0	0.2	1.0	1.1
N^1	0.0	0.3	0.0	3.4
N^2	1.0	0.2	0.0	0.5

(a) Reactants formulas and reference table R_1

	Elements σ			
	C	O	N	H
C^1	0.8	0.5	0.9	2.3
C^2	0.8	2.0	0.1	0.6
N^1	1.0	0.3	0.0	2.6
N^2	0.0	0.3	0.0	3.4

(b) Products formulas and reference table R_2

Figure 5: Reactants and products with reference coordination matrices used to define s_2 and z_2 for metadynamics. The choice of atoms' numeral is mapped on the Strecker pathway (See supporting information **S-1-2**).

1. Metadynamics is used to explore the possible intermediates for the reaction, and find a continuous pathway from reactants to products.[42]

2. Transition Path Sampling (TPS) methods are used to characterize every transition of every elementary steps,[43–45] and build appropriate CVs.

3. Umbrella Sampling (US) is then used to estimate the Free Energy Profile (FEP) of every steps.[46]

This procedure organisation ensure the agnosticism of this protocol making it reliable and autonomous regarding potentially biased choices, leaving to the researcher the control of the quality of the results.

The *ab initio* calculations are performed by the mean of density functional theory (DFT), using the PBE exchange and correlation functionals[47] and D2 Grimme's van der Waals empirical corrections[48] along with Martins-Troullier pseudopotentials,[49] and a plane wave

basis set with a cutoff of 80 Ry. The convergence cutoff for orbital optimization has been fixed to 10^{-5}. This setup has been shown to describe satisfactorily the electronic structure in our previous works.[12,13,50] We checked the quality of DFT-PBE predictions by carrying out single point potential energy calculations with the hybrid PBE0 functional. Details and an initial evaluation are in the Supplementary Information (**S-5**). We estimate a potential energy difference between PBE and PBE0 of about 5-7 kcal/mol for transition states and 2-3 kcal/mol for intermediates.

All simulations were performed at 300K in the NVT ensemble, using Nose-Hoover chained thermostats[51,52] and a time step of 20 a.u. (≈0.48 fs). The system consists in a periodically repeated cubic unit cell of 13.36Å. The reacting molecules of the synthesis are one formaldehyde, one hydrogen cyanide and two ammonia. They are solvated with 80 explicit water molecules. The total represents 251 atoms (designed as 'N' in the following). The size of the box has been fixed as to correspond to a density of 1.058 kg/L. The concentration of solvated compounds is therefore ≈0.7 mol/L. The temperature at 300K and density at 1.058 kg.L^{-1} have been set this way in order to compare the results with preceding literature.[5,22,24,39]

Our CPMD input template for AIMD simulations can be found in supporting information (See **S-1-1**). All input details are freely available on PLUMED-NEST (www.plumed-nest.org), the public repository of the PLUMED consortium.[41] All the essential parameters of the protocol steps can be found at the end of the supporting information (See **S-6**).

The 3·N dimensional atomic position space, also known as configurational space, poses enormous challenges in terms of efficient sampling, or even for visualization. To address this issue, it is crucial to reduce the number of variables to a very limited set. This reduction must enable an effective monitoring of transformations in this highly-dimensional space, and a robust understanding of the essential degrees of freedom involved in those transformations. These condensed variables are referred to as Collective Variables (CVs).

In particular, Path Collective Variables consist of two variables, denoted as s, z, as introduced by Branduardi *et al*[53] in 2007. In our chemical system we derive them from an

ensemble of reference coordination matrices, represented as R_i, as in F. Pietrucci and A. M. Saitta paper[54] in 2015.

In particular, a coordination matrix is constructed using a specific set of atoms within the system, corresponding to a particular geometry. This matrix contains the coordination numbers of each atom in the set (rows) for every element present in the periodic table (columns). The initial ensemble of reference coordination matrices for our protocol, illustrating the reactants and products of the complete mechanism, is shown in Figure 5.

Typically, the variables s, z are assigned indices of the closest reference matrix, exhibiting complementary behaviors:

⇒ The variable s indicates the index of the closest reference matrix. In our initial set (Figure 5), if the system nears the reactive reference matrix, s_2 approaches 1, and if closer to the products, it nears 2. Throughout the protocol, s serves as the presumed reaction coordinate.

⇒ The variable z indicates the topological distance from the complete reference set. A z value over 1 indicates the sampling of geometries not covered by the references. z provides critical information about system behavior during simulations, especially in quantitative analyses.

The s, z set has proven to be a powerful tool for tracking reaction mechanisms in solution using enhanced sampling techniques.[13,14,53,54]

Metadynamics

To address the slow reactivity of chemical systems within the limits of typical simulation times, we adopt metadynamics. This approach involves forcing the system to progress from reactants to products by systematically incrementing an external potential, added to the DFT potential, during the dynamics.[42,55] This potential takes the form of a sum of Gaussian functions, acting on both $s(t)$ and $z(t)$, facilitating a large exploration of reaction pathways.

For this step, we employ only two reference matrices presented in Figure 5). The functions s and z are then denoted as s_2 and z_2, respectively. Following the metadynamics, if stable intermediate geometries are identified, they are incorporated into the reaction mechanism. Ultimately, the pathway is composed only of successive elementary steps.

Committor Analysis (CA)

Committor Analysis (CA)[56] estimates the probability of transitioning into reactants rather than products when starting a molecular dynamics simulation from a given geometry with initial velocities drawn randomly from the Maxwell-Boltzmann distribution. The committor value helps identify transition states, when its value is 0.5.

To identify transition states from metadynamics trajectories, we sample geometries at intervals of 10 steps (approximately 5 fs) and perform CA. 20 trajectories are generated from each geometry, and the one with an estimated committor value closest to 0.5 is chosen. These trajectories form a dataset spanning 10 ps (around 20,000 points), which is used to derive data-driven reaction coordinates.

For the final elementary step, CA alone did not yield satisfactory transitions for subsequent protocol stages. Hence, we employ the Shooting From the Top (SFT) approach,[57] a transition path sampling method aiming to capture a reactive trajectory close to the Minimum Free Energy Path (MFEP) specifically for this step. This refined trajectory undergoes CA again to generate a higher-quality dataset. Details of the SFT process are provided in Supplementary Information (See **S-1-5**).

Highly referenced PCV

In order to catch the fine details of the mechanism in the final quantitative step, we generated two new PCVs for each step. These PCVs were constructed using 12 internal references denoted as R_i. To obtain these references, we carried out an exhaustive search across all potential matrices within our dataset derived from CA, using the reference space exploration

algorithm described in our previous publication.[14] Additional technical details can be found in the Supplementary Information (See **S-1-6**).

For chemical steps more complex than deprotonation, we used highly referenced PCVs in umbrella sampling (US). However, since deprotonation is a relatively straightforward process, we opted to use the proton coordination of the relevant atom as the Reaction Coordinate (RC).

Umbrella Sampling (US)

Umbrella Sampling[46,58] constitutes the quantification step in the protocol designed to derive the Free Energy Profile (FEP) associated with each reaction step. A more detailed description of this process is available in the Supplementary Information (See **S-1-4**).

This phase of the protocol is the most resource-intensive, demanding a combined 0.9 ns of trajectories, equivalent to 360,000 CPU hours, for a single full *ab initio* calculation. Additionally, it involve a meticulous verification of result quality due to the sensitivity of the process. To calculate the FEP we use the Weighted Histogram Analysis Method (WHAM), using the code developed by A. Grossfield.[59]

To overcome the intrinsic computational limitations of AIMD, we use Machine Learning Interatomic Potentials (MLIPs), trained on a select set of umbrella sampling windows, then allowing the use of *ab initio*-quality Machine Learning Molecular Dynamics (MLMD). The identification of specific simulations for training, as well as the construction of the training set, follows the methodology described in reference.[15]

For each elementary step in the mechanism, four MLIPs are trained using the DeePMD-kit smooth edition[60,61] with distinct random seeds. This ensemble of MLIPs functions as a committee, and the assessment of the maximum deviation in force predictions among committee members allows for continuous monitoring of prediction quality.

More details for the MLIP procedure, including a table with a comparison of relative computational costs can be found in the Supplementary Information (See subsection **S-1-7**).

Using MLMD results in a sixfold cost reduction of Umbrella Sampling (US), the most expensive step of the protocol. A typical cost evaluation is provided in the supporting information (see Table **S-4**). This efficiency gain is expected to be even more pronounced for larger and more complex systems, making quantum-accuracy free-energy calculations more accessible. This advancement enables the exploration of systems closer to experimental conditions, including additional solvent molecules or surface-activated reactions.

Acknowledgement

We thank F. Siro Brigiano and F. Pietrucci for critical reading of the article. We especially thanks F. Pietrucci for PLUMED modified codes, theoretical support and precious advice. We also thank A. France-Lanord and R. Vuilleumier for helpful discussions. This work has been supported by the institute of computing and data sciences (ISCD) of Sorbonne University which funds L. Huet's thesis.

This work was performed thanks to High Performance Computing (HPC) resources from GENCI (Grand Equipement National de Calcul Intensif), on Jean-Zay cluster of Idris (Institut du Développement et des Ressources en Informatique Scientifique) (Grants 2024-A0160901387, 2023-A0140901387, 2022-A0120901387, and 2021-A0100901387).

Supporting Information Available

Supplementary information about the method details and the data checking are freely available in a separated file.

References

(1) Pizzarello, S.; Weber, A. L. Prebiotic Amino Acids as Asymmetric Catalysts. *Science* **2004**, *303*, 1151–1151.

(2) Elsila, J. E.; Glavin, D. P.; Dworkin, J. P. Cometary glycine detected in samples returned by Stardust. *Meteoritics & Planetary Science* **2009**, *44*, 1323–1330, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1945-5100.2009.tb01224.x.

(3) Sandford, S. A. *et al.* Organics Captured from Comet 81P/Wild 2 by the Stardust Spacecraft. *Science* **2006**, *314*, 1720–1724, Publisher: American Association for the Advancement of Science.

(4) Snyder, L. E.; Lovas, F. J.; Hollis, J. M.; Friedel, D. N.; Jewell, P. R.; Remijan, A.; Ilyushin, V. V.; Alekseev, E. A.; Dyubko, S. F. A Rigorous Attempt to Verify Interstellar Glycine. *The Astrophysical Journal* **2005**, *619*, 914, Publisher: IOP Publishing.

(5) Miller, S. L.; Van Trump, J. E. The Strecker Synthesis in the Primitive Ocean. Origin of Life. Dordrecht, 1981; pp 135–141.

(6) Burton, A. S.; Stern, J. C.; Elsila, J. E.; Glavin, D. P.; Dworkin, J. P. Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chemical Society Reviews* **2012**, *41*, 5459–5472, Publisher: The Royal Society of Chemistry.

(7) Pizzarello, S.; Shock, E. The Organic Composition of Carbonaceous Meteorites: The Evolutionary Story Ahead of Biochemistry. *Cold Spring Harbor Perspectives in Biology* **2010**, *2*, a002105, Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

(8) Ehrenfreund, P.; Charnley, S. B. Organic Molecules in the Interstellar Medium, Comets, and Meteorites: A Voyage from Dark Clouds to the Early Earth. *Annual Review of Astronomy and Astrophysics* **2000**, *38*, 427–483, _eprint: https://doi.org/10.1146/annurev.astro.38.1.427.

(9) Miller, S. L. Production of Some Organic Compounds under Possible Primitive Earth Conditions [1]. *Journal of the American Chemical Society* **1955**, *77*, 2351–2361.

(10) Koga, T.; Naraoka, H. A new family of extraterrestrial amino acids in the Murchison meteorite. *Scientific Reports* **2017**, *7*, 636, Number: 1 Publisher: Nature Publishing Group.

(11) Krishnamurthy, R. Life's Biological Chemistry: A Destiny or Destination Starting from Prebiotic Chemistry? *Chemistry A European Journal* **2018**, *24*, 16708–16715, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/chem.201801847.

(12) Pérez-Villa, A.; Pietrucci, F.; Saitta, A. M. Prebiotic chemistry and origins of life research with atomistic computer simulations. *Physics of Life Reviews* **2020**, *34-35*, 105–135.

(13) Magrino, T.; Pietrucci, F.; Saitta, A. M. Step by Step Strecker Amino Acid Synthesis from Ab Initio Prebiotic Chemistry. *The Journal of Physical Chemistry Letters* **2021**, *12*, 2630–2637, Publisher: American Chemical Society.

(14) Magrino, T.; Huet, L.; Saitta, A. M.; Pietrucci, F. Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry. *The Journal of Physical Chemistry A* **2022**, *126*, 8887–8900, Publisher: American Chemical Society.

(15) Devergne, T.; Magrino, T.; Pietrucci, F.; Saitta, A. M. Combining Machine Learning Approaches and Accurate Ab Initio Enhanced Sampling Methods for Prebiotic Chemical Reactions in Solution. *Journal of Chemical Theory and Computation* **2022**, *18*, 5410–5421, Publisher: American Chemical Society.

(16) Benayad, Z.; David, R.; Stirnemann, G. Prebiotic chemical reactivity in solution with quantum accuracy and microsecond sampling using neural network potentials. *Proceedings of the National Academy of Sciences* **2024**, *121*, e2322040121, Publisher: Proceedings of the National Academy of Sciences.

(17) Kortum, G.; Vogel, W.; Andrussow, K. *DISSOCIATION CONSTANTS OF ORGANIC ACIDS IN AQUEOUS SOLUTION*.

(18) Schlesinger, G.; Miller, S. L. Equilibrium and kinetics of glyconitrile formation in aqueous solution. *Journal of the American Chemical Society* **1973**, *95*, 3729–3735, Publisher: American Chemical Society.

(19) Etim, E. E. Benchmark Studies on the Isomerization Enthalpies for Interstellar Molecular Species. 2023; `http://arxiv.org/abs/2302.05911`, arXiv:2302.05911 [astro-ph].

(20) Jammot, J.; Pascal, R.; Commeyras, A. The influence of borate buffers on the hydration rate of cyanohydrins: evidence for an intramolecular mechanism. *Journal of the Chemical Society, Perkin Transactions 2* **1990**, *1*, 157–162, Publisher: The Royal Society of Chemistry.

(21) Jammot, J.; Pascal, R.; Commeyras, A. Hydration of cyanohydrins in weakly alkaline solutions of boric acid salts. *Tetrahedron Letters* **1989**, *30*, 563–564.

(22) Haberfield, P. What is the energy difference between H2NCH2CO2H and +H3NCH2CO2-? *Journal of Chemical Education* **1980**, *57*, 346, Publisher: American Chemical Society.

(23) Smirnov, V. I.; Badelin, V. G. The enthalpies of solution and solvation of glycine in mixed water-formamide solvents at 298.15 K. *Russian Journal of Physical Chemistry* **2006**, *80*, 357–360.

(24) Thrush, K. L.; Kua, J. Reactions of Glycolonitrile with Ammonia and Water: A Free Energy Map. *The Journal of Physical Chemistry A* **2018**, *122*, 6769–6779, Publisher: American Chemical Society.

(25) Riley, K. E.; Op't Holt, B. T.; Merz, K. M. Critical assessment of the performance

of density functional methods for several atomic and molecular properties. *Journal of chemical theory and computation* **2007**, *3*, 407–433.

(26) Arnaud, R.; Adamo, C.; Cossi, M.; Milet, A.; Vallée, Y.; Barone, V. Theoretical Study of the Addition of Hydrogen Cyanide to Methanimine in the Gas Phase and in Aqueous Solution. *Journal of the American Chemical Society* **2000**, *122*, 324–330, Publisher: American Chemical Society.

(27) Magrino, T.; Pietrucci, F.; Saitta, A. M. Correction to: Step by Step Strecker Amino Acid Synthesis from ab Initio Prebiotic Chemistry (submitted). *The Journal of Physical Chemistry Letters* **2024**, -.

(28) Moutou, G.; Taillades, J.; Bénefice-Malouet, S.; Commeyras, A.; Messina, G.; Mansani, R. Equilibrium of a-aminoacetonitrile formation from formaldehyde, hydrogen cyanide and ammonia in aqueous solution: Industrial and prebiotic significance. *Journal of Physical Organic Chemistry* **1995**, *8*, 721–730, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/poc.610081105.

(29) Stribling, R.; Miller, S. L. Energy yields for hydrogen cyanide and formaldehyde syntheses: The hcn and amino acid concentrations in the primitive ocean. *Origins of life and evolution of the biosphere* **1987**, *17*, 261–273.

(30) Cooper, G.; Kimmich, N.; Belisle, W.; Sarinana, J.; Brabham, K.; Garrel, L. Carbonaceous meteorites as a source of sugar-related organic compounds for the early Earth. *Nature* **2001**, *414*, 879–883, Number: 6866 Publisher: Nature Publishing Group.

(31) Cronin, J. R.; Chang, S. In *The Chemistry of Life's Origins*; Greenberg, J. M., Mendoza-Gómez, C. X., Pirronello, V., Eds.; NATO ASI Series; Springer Netherlands: Dordrecht, 1993; pp 209–258.

(32) Ceselin, G.; Salta, Z.; Bloino, J.; Tasinato, N.; Barone, V. Accurate Quantum Chemical Spectroscopic Characterization of Glycolic Acid: A Route Toward its Astrophysical

Detection. *The Journal of Physical Chemistry A* **2022**, *126*, 2373–2387, Publisher: American Chemical Society.

(33) Mardyukov, A.; Keul, F.; Schreiner, P. R. 1,1,2-Ethenetriol: The Enol of Glycolic Acid, a High-Energy Prebiotic Molecule. *Angewandte Chemie* **2021**, *133*, 15441–15444, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ange.202104436.

(34) Nuevo, M.; Bredehöft, J. H.; Meierhenrich, U. J.; d'Hendecourt, L.; Thiemann, W. H.-P. Urea, Glycolic Acid, and Glycerol in an Organic Residue Produced by Ultraviolet Irradiation of Interstellar/Pre-Cometary Ice Analogs. *Astrobiology* **2010**, *10*, 245–256, Publisher: Mary Ann Liebert, Inc., publishers.

(35) Taillades, J.; Beuzelin, I.; Garrel, L.; Tabacik, V.; Bied, C.; Commeyras, A. N-Carbamoyl-α-Amino Acids Rather than Free α-Amino Acids Formation in the Primitive Hydrosphere: A Novel Proposal for the Emergence of Prebiotic Peptides. *Origins of life and evolution of the biosphere* **1998**, *28*, 61–77.

(36) Pietrucci, F.; Aponte, J. C.; Starr, R.; Pérez-Villa, A.; Elsila, J. E.; Dworkin, J. P.; Saitta, A. M. Hydrothermal decomposition of amino acids and origins of prebiotic meteoritic organic compounds. *ACS Earth and Space Chemistry* **2018**, *2*, 588–598.

(37) Klingler, D.; Berg, J.; Vogel, H. Hydrothermal reactions of alanine and glycine in sub- and supercritical water. *The Journal of Supercritical Fluids* **2007**, *43*, 112–119.

(38) Sato, N.; Quitain, A. T.; Kang, K.; Daimon, H.; Fujie, K. Reaction Kinetics of Amino Acid Decomposition in High-Temperature and High-Pressure Water. *Industrial & Engineering Chemistry Research* **2004**, *43*, 3217–3222, Publisher: American Chemical Society.

(39) Danger, G.; Duvernay, F.; Theulé, P.; Borget, F.; Chiavassa, T. HYDROXYACETONITRILE (HOCH2CN) FORMATION IN ASTROPHYSICAL CONDITIONS. COMPE-

TITION WITH THE AMINOMETHANOL, A GLYCINE PRECURSOR. *The Astrophysical Journal* **2012**, *756*, 11, Publisher: The American Astronomical Society.

(40) Hutter, J.; Curioni, A. Car–Parrinello Molecular Dynamics on Massively Parallel Computers. *ChemPhysChem* **2005**, *6*, 1788–1793, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.200500059.

(41) Bonomi, M. *et al.* Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods* **2019**, *16*, 670–673, Number: 8 Publisher: Nature Publishing Group.

(42) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *WIREs Computational Molecular Science* **2011**, *1*, 826–843, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.31.

(43) Bolhuis, P. G.; Swenson, D. W. Transition Path Sampling as Markov Chain Monte Carlo of Trajectories: Recent Algorithms, Software, Applications, and Future Outlook. *Advanced Theory and Simulations* **2021**, *4*, 2000237.

(44) Dellago, C.; Bolhuis, P. G. Transition path sampling and other advanced simulation techniques for rare events. *Advanced computer simulation approaches for soft matter sciences III* **2009**, *221*, 167–233.

(45) Juraszek, J.; Vreede, J.; Bolhuis, P. G. Transition path sampling of protein conformational changes. *Chemical Physics* **2012**, *396*, 30–44.

(46) Kästner, J. Umbrella sampling. *WIREs Computational Molecular Science* **2011**, *1*, 932–942, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.66.

(47) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996**, *77*, 3865–3868, Publisher: American Physical Society.

(48) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006**, *27*, 1787–1799, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20495.

(49) Troullier, N.; Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Physical Review B* **1991**, *43*, 1993–2006, Publisher: American Physical Society.

(50) Saitta, A. M.; Saija, F. Miller experiments in atomistic computer simulations. *Proceedings of the National Academy of Sciences* **2014**, *111*, 13768–13773.

(51) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **1984**, *81*, 511–519.

(52) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **1985**, *31*, 1695–1697, Publisher: American Physical Society.

(53) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *The Journal of chemical physics* **2007**, *126*, 054103.

(54) Pietrucci, F.; Saitta, A. M. Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios. *Proceedings of the National Academy of Sciences* **2015**, *112*, 15030–15035, Publisher: Proceedings of the National Academy of Sciences.

(55) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(56) Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annual Review of Physical Chemistry* **2016**, *67*, 669–690, _eprint: https://doi.org/10.1146/annurev-physchem-040215-112215.

(57) Jung, H.; Okazaki, K.-i.; Hummer, G. Transition path sampling of rare events by shooting from the top. *The Journal of chemical physics* **2017**, *147*, 152716.

(58) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.

(59) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.11. `http://membrane.urmc.rochester.edu/?page_id=126`, Last accessed 28 September 2022.

(60) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications* **2018**, *228*, 178–184.

(61) Zhang, L.; Han, J.; Wang, H.; Saidi, W. A.; Car, R.; E, W. End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems. *arXiv:1805.09003 [cond-mat, physics:physics]* **2018**, *31*, arXiv: 1805.09003.

5.2 Conclusion in the context of the Thesis

In this paper, we proved the robustness of our method, as a thorough review of the two measures of the glycine synthesis free energy gaps finally showed good agreement considering our statistical error bars.

The results presented in this paper now need to be shared with the scientific community. For other theoretical researchers, it represents a new step towards the use of machine learning interatomic potentials for chemical reactions in solution and in prebiotic chemistry. From an experimental point of view, it provides thermodynamic results to support experimentation along with new proposed intermediates to investigate the synthesis of glycine.

Chapter 6

Exploration of the dynamics of a S_N2

6.1 Presentation of the article

The flowing article is about to be submitted for publication. Despite the calculations being finished, it is only in the pre-print stage and still needs a proofreading. The main idea was to incorporate an ultimate step in our protocol to infer the kinetics of our studied pathway. To do so, we wanted to use insights from advanced stochastic equations developed in recent years [107, 96, 98]. However, the possibility of applying these tools in the concrete case of chemical reaction in solution was still to be proven.

In this article, we give a description of the challenges ahead and point out the critical points of our methodologies that remain to be addressed to apply these tools in a fully agnostic framework. We performed a critical reading of the methodologies and test various methods for inference of the kinetics in order to identify the most reliable and affordable ones.

The supporting information for this article is available in the Appendix E.

Insight on chemical reaction dynamics and reaction coordinates from non-Markovian models

Léon Huet,[†] Hadrien Vroylandt,[‡] Rodolphe Vuilleumier,[¶,§] A. Marco Saitta,[†] and Fabio Pietrucci[*,†]

†*Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS UMR 7590, Paris 75005 FRANCE*

‡*Institut des sciences du calcul et des données, Sorbonne Université, 75005 Paris, France*

¶*PASTEUR, Département de Chimie, Ecole Normale Supérieur, PSL University, Sorbonne Université, CNRS, 75005 Paris, France*

§

E-mail: fabio.pietrucci@sorbonne-universite.fr

Abstract

We present a comprehensive analysis of the dynamic behavior of reaction coordinates using the generalized Langevin equation framework. Our investigation focuses on identifying quality criteria for collective variables to accurately describe the chemical reactions, with particular emphasis on their dynamic properties and memory effects. Using the S_N2 reaction of methyl chloride with a chloride ion as a model system, we demonstrate that traditional methods, which primarily focus on static properties such as free energy barriers, are insufficient for dynamic studies. We systematically examine the correlation between memory kernels, friction coefficients, transmission coefficients, and the committor, revealing that criteria obtained from the friction in the well ensemble are not correlated with the behavior of the CVs at the top of the barrier.

Moreover, our findings highlight the necessity of a smooth and deterministic effective mass curve for CVs, which is crucial for reliable dynamic inference. The study further underscores the limitations of data-driven CVs, which often involve complex dependencies on multiple degrees of freedom, complicating their use in dynamic modeling. This approach emphasizes the importance of considering both static and dynamic properties in the development of effective collective variable for the accurate inference of reaction kinetics.

Using a heuristic reliable reaction coordinate, and by combining Umbrella Sampling results for free energy estimation with well equilibrium data for friction estimation, we have constructed a GLE model that yields reaction rates in reasonable agreement with experimental data, after rescaling the barrier heights.

Introduction

In the theoretical field of chemical reactions in solution, the developed tools offer results with errors of up to two orders of magnitude concerning the kinetics.[1,2] This is mainly due to two factors: the kinetic rate evolves exponentially with the activation barrier of the mechanism,[3] and the high barriers prevent a true understanding of the dynamics of these systems, making it difficult to sample the reaction process correctly. These factors make the development of accurate tools for inferring the kinetics of chemical reactions a complex issue.

Chemical reactions, particularly those involving covalent bonding and cleaving, require overcoming highly energetic transition states.[4] A typical barrier height is around 20 kcal/mol (33.5 k_BT units at 300 K), corresponding to a transition timescale on the order of minutes, making it impractical to wait for spontaneous transitions, especially as *ab initio* level force estimation is required. Even with affordable simulation times, the disparity and complementarity of the methods used for rate estimation make it difficult to determine which one provides the best results,[1] particularly when an accurate mean first passage time (MFPT) estimation is not feasible.

Additionally, choosing an appropriate collective variable (CV) is essential for kinetic inference for two reasons. It influences the quality of barrier height and friction estimation—two crucial factors in the dynamics of the chemical process in solution. It may also correctly separate the irrelevant degrees of freedom of the solvent from those mandatory for the reaction[4,5] which helps us to understand the mechanism.

Historically, the tools used to estimate the kinetics of chemical processes are the stochastic differential equations.[6,7] Their use separates the solvent degrees of freedom into a noise and serves as a theoretical basis for developing analytical tools for rate estimation.[1,3]

In this paper, we model the effective dynamics of CVs. This method is challenging as it differs significantly from the direct estimation of the dynamics of a particle in suspension, especially concerning the mass of the studied phenomena, which becomes an effective mass. We turn to Langevin models of the dynamics. Several variants of Langevin equations are available, and in this work, we focus on the Generalized Langevin Equation (GLE) and Underdamped Langevin Equation (ULE).[7–9] Using these tools, we study the kinetic behavior of a paradigmatic chemical reaction in explicit solution and explore various methods to calculate its rate.[1,10]

We perform a complete *ab initio* Molecular Dynamics (AIMD) study of the reaction to obtain DFT-PBE level data of the well and the transition path and to estimate the barrier of the transition. With the collected data, we conduct a comprehensive analysis of different indicators of the dynamic properties of the system, such as memory effects and friction associated with a CV, as well as the correlation between the CVs and the committor.[4,11]

We also discuss the possibility of inferring a non-Markovian dynamic model based on the GLE from these types of systems and using it to estimate their kinetics.

Methods

The system: Methyl-chloride and chloride ion

Our studied system contains 98 water molecules, one chloride ion, one methyl chloride molecule, and one potassium counter-ion. It is labeled MCCI, for Methyl Chloride and Chloride Ion, and consists of a cubic box of 14.4807 Å, employing periodic boundary conditions. The studied reaction is the S_N2 reaction, which involves the chloride ion as the nucleophile and the carbon of the methyl chloride as the electrophile. This reaction has been studied in many experimental and theoretical papers.[12–17] Due to the permutation invariance of the two chlorides, this reaction has equivalent reactants and products. This guarantees that the reaction free energy, internal energy, and specific entropy are all zero. This implies that the only relevant feature of this reaction is its kinetics.

The Born-Oppenheimer molecular dynamics of the full system are conducted using the CPMD software.[18] All atoms are treated at the same theoretical level with DFT. The chosen exchange-correlation functional is PBE[19] with D2 Grimme empirical correction.[20] The wavefunctions are expanded using a plane wave basis set with a cutoff of 80 Ry. Core electrons are represented using Martins-Troullier pseudopotentials.[21] The temperature is fixed with a Nose-Hoover chain thermostat[22–24] at 300 K. Our CPMD input template can be found in the Supplementary Information (See **S-1**).

Due to the high transition barrier associated with chemical systems, AIMD permits a sampling of the configuration space that is limited in quantity and quality. These limitations result in different datasets with distinct properties, corresponding to various sampling methods. In this paper, we generate three kinds of data:

1. AIMD data without bias (or with a static bias assimilated as the background potential) in the reactant and product wells of the reaction. Since the reaction barrier is high, AIMD within these wells will not cross it within a reasonable simulation time. This allows for obtaining unbiased data of the system that reaches equilibrium, but only in

the vicinity of reactants or products.

2. Transition Path Sampling (TPS) data, which can be obtained using various methods.[25] In this study, our TPS dataset consists of 600 short trajectories (1.5 ps) initiated from the top of the barrier with a Maxwell-Boltzmann distribution of velocities, without additional external forces compared to the AIMD dataset. The protocol used to generate this dataset is presented in our previous article.[4]

3. Umbrella Sampling (US) data, comprising biased trajectories confined to specific regions of a reaction coordinate (RC) by an external harmonic bias.[26] Each trajectory in the US dataset overlaps with adjacent trajectories to cover all possible values of the RC. This is achieved by placing the harmonic bias at regular intervals along the RC for each trajectory. This dataset is composed exclusively of trajectories that have reached a static state thanks to this external constraint. Specifically, US data can be employ to determine the Helmholtz free energy profile of the RC using the WHAM algorithm.[27]

Model

The two models used in this paper are derived from the GLE. This generic model has proven to be exact in the case of position-dependent memory for any CVs within the configuration space Γ, which denotes the phase space positions $\{q\}$.[9,28] Since the dataset required to infer the position-dependent memory kernel for a reaction in solution remains unattainable, we focus on the position-independent version of the memory. A discussion regarding this choice is provided in the discussion section. In this context, the GLE is expressed as:

$$\begin{aligned} &\forall x \text{ a CV} \quad ; \quad \dot{x} = v \\ &\dot{v} = f(x(t)) - \int_0^\infty K(\tau) v(t-\tau)\, d\tau + \xi(t) \end{aligned} \tag{1}$$

In Equation 1, f designates the effective mean force, ξ denotes the noise, which will be

detailed subsequently.

In the case we consider, the phase space evolution follows a Hamiltonian in the (N,V,T) ensemble. We define the marginal Helmholtz free energy F associated with x as:

$$F(x) = -k_B T \ln(\rho_{x,\mathrm{eq}}(x)) \tag{2}$$

Additionally, we estimate the effective mass of the variable x, denoted m_{eff}, using two different methods. These methods are connected through the Maxwell-Boltzmann distribution of velocities at equilibrium:

$$m_{eff}(x)^{-1} = \frac{1}{k_B T}\langle v^2 | x\rangle \tag{3}$$

$$= \langle \nabla_{\{q\}}(x) M^{-1} \nabla_{\{q\}}(x) | x\rangle \tag{4}$$

where M is the diagonal matrix containing the masses of the elementary particles in the system. The use of Equation 3 to infer the effective mass is applicable only to data where stationarity has been achieved.

From F and m_{eff}, we can detail the composition of f and K in Equation 1, assuming that the effective mass is also independent of the position x:

$$f(x) = -m_{eff}^{-1} \nabla_x F(x) \tag{5}$$

$$K(\tau) = \mathbf{K}(\tau) m_{eff}^{-1} \tag{6}$$

A crucial step in estimating a GLE model is the reconstruction of the memory kernel. Several approaches have been proposed in the past; however, they are limited to cases where ergodic MD trajectories serve as input data, i.e., repeatedly sampling transitions between

all the meta-stable states.[2,9,29–31]

A straightforward approach involves the numerical solution of the Volterra equation:

$$\left\langle v(0)\left(\dot{v}(\tau) - f(x(\tau))\right)\right\rangle = -\int_{-\infty}^{0} dt' K(-t')\langle v(0)v(\tau+t')\rangle \tag{7}$$

To invert this equation, we use the method implemented in an open-source code,[32] which is based on the trapezoidal method developed by P. Linz in 1969.[33]

GLE models are expected to faithfully reproduce the average behavior of the real system down to the femtosecond time resolution Δt. However, it is possible,in principle, to estimate an underdamped model when considering a time resolution coarser than the decay time of the memory kernel:

$$\dot{v}(t) = f(x(t)) - \gamma v(t) + \xi(t) \tag{8}$$

A first estimation of the friction coefficient γ can be obtained employing the integral of the memory kernel. We denote the friction coefficient derived from this definition as γ_0:

$$\gamma_0 = \int_0^{\infty} K(\tau) \tag{9}$$

The reliability of γ_0 in Equation 9 is compromised by a potentially long memory kernel, which may be comparable to or exceed the characteristic decorrelation time of the velocity.

An alternative method for inferring the underdamped friction coefficient from the GLE memory kernel involves calculating the linear correlation coefficient between the instantaneous velocity $v(t)$ and the GLE instantaneous friction term, as discussed in the work of D. Girardier.[34] This coefficient estimation, using the linear model, is denoted as γ_1:

$$\gamma_1 = \langle \frac{\int_0^{t_0} K(\tau)v(t-\tau)d\tau}{v(t)} \rangle \tag{10}$$

where t_0 represents the upper limit of integration of the memory kernel.

We can also introduce R_1^2, the coefficient of determination resulting from the estimation of γ_1, as a criterion for assessing the degree of Markovian behavior in the dynamics of x.

Another estimation method, as discussed in the same reference, involves using the autocorrelation of the velocity $C_v(\tau)$ to calculate the resulting friction coefficient γ_2, as defined in Equation 11. γ_2 is determined by fitting the exponential decay of $C_v(\tau)$ according to the following equation:

$$C_v(\tau) = \frac{\langle v(0)v(\tau)\rangle}{\langle v^2\rangle} \equiv e^{-\frac{\gamma_2\tau}{2}}\Omega(\tau) \tag{11}$$

where $\Omega(\tau)$ is a periodic function. This approach depends on the accuracy of the fit, which becomes increasingly challenging in non-harmonic potentials typical of chemical systems.

A final approach is to solve the Markovian version of the Volterra equation 7, where the memory kernel takes the form of a Dirac delta function, known as the Kramers-Moyal equation.

$$\gamma_3 = \frac{\langle v(0)\left(\dot{v}(\tau) - f(x(\tau))\right)\rangle}{\langle v(0)v(\tau)\rangle} \tag{12}$$

We will evaluate these three last definitions of the friction coefficient to determine which is the most suitable for constructing an underdamped dynamic model of the reaction.

Collective variables

Heuristic Variable

The standard heuristic RC used for the MCCI system is Δd, defined as the difference in distance between d_1 and d_2, the distances from the carbon of the methyl group to the two chloride nuclei. In our previous study, we demonstrated that this variable serves as an effective RC.[4]

To estimate the kinetics of a reaction, a crucial component is the free energy curve

associated with the selected RC. Given that the US simulation is both necessary and resource-intensive for this measurement, we applied it to the most promising candidate. Consequently, this study focuses on efficient sampling of the configuration space with respect to Δd.

The Δd region examined was confined to the interval $[-4, 4]$ Bohr, with two additional quadratic walls incorporated into the background potential to facilitate rapid stabilization within the well ensemble. While these additional walls affect the system's kinetics, crossing the barrier remains the primary limiting process, and thus, the impact of these walls should be minimal. Further investigation into the solvation of the anion using a classical potential could complement this study.

From the definition of Δd and equation 4, we can derive a purely analytical formula for the effective mass of this variable:

$$\frac{1}{m_{\Delta \mathrm{d}}} = \langle \frac{2}{m_{Cl}} + \frac{2}{m_C}(1 - \cos(\theta)) \rangle \tag{13}$$

where θ is the angle $\widehat{Cl_1, C, Cl_2}$.

As observed, even for a simple linear combination of two distances, the definition of the instantaneous mass of this CV depends on other degrees of freedom, specifically the angle θ.

We estimate the mass using two different formulas: one from equation 3, which is applicable to all variables, and the other from equation 13. Our initial estimation of the mass for the system in the reactant well is provided in the Supplementary Information (See **S-2**). This initial analysis revealed that the variable θ required a significant amount of time to stabilize within the well ensemble. Even after an unbiased trajectory of 100 ps, stabilization was incomplete. This slow stabilization is attributed to numerous solvation metastable states that correspond to various values of θ but remain within the $[-4, 4]$ Bohr interval of Δd. To address this issue, we repeated our initial dynamics simulation by adding a new external upper wall to the background potential specifically for the $\cos(\theta)$ CV at -0.7.

The final working intervals for this study are thus $[-4, 4]$ for Δd (in Bohr) and $[-1, -0.7]$

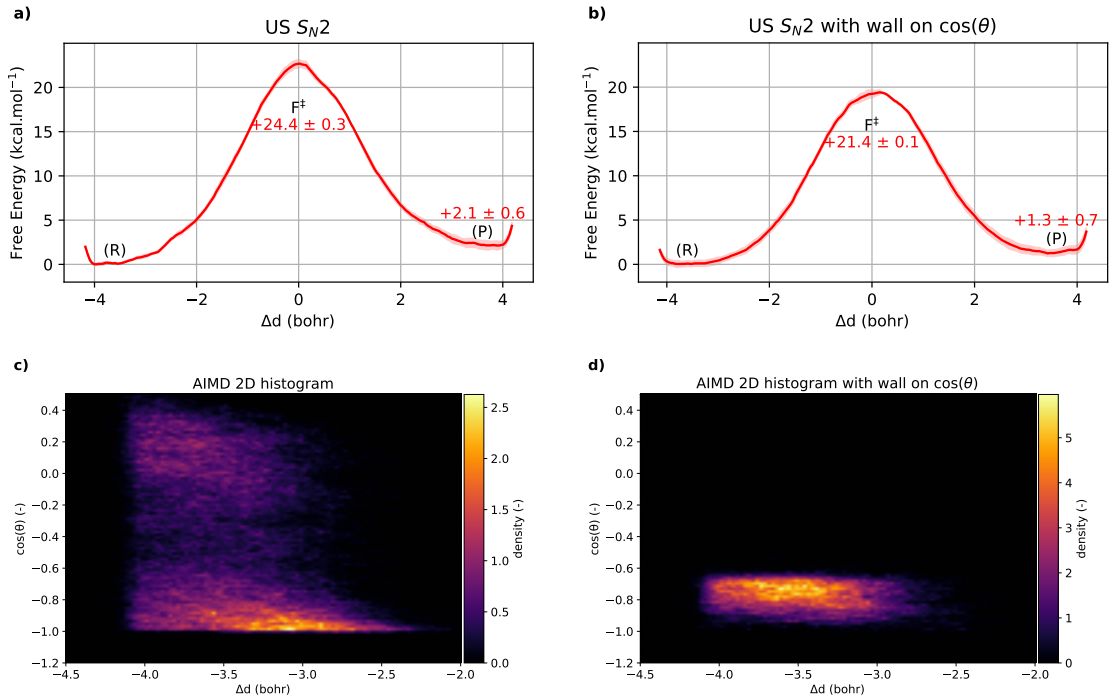


Figure 1: Impact of the added wall on $\cos(\theta)$ on the free energy profile and the behavior of the AIMD dataset. **a)** Free energy profile of the reaction with restrictive walls applied only to Δd. **b)** Same as panel **a)**, but with an additional wall on $\cos(\theta)$. The lower barrier results from restricted exploration within the well ensemble. **c)** 2D histogram of the (Δd, $\cos(\theta)$) pair without a wall on $\cos(\theta)$. The slow diffusion of $\cos(\theta)$ values indicates non-stationarity of the histogram. **d)** Same as panel **c)**, but with the additional wall on $\cos(\theta)$. In this case, stationarity of both variables is achieved.

for $\cos(\theta)$. The transition state ensemble is close to 0 for Δd and -1 for $\cos(\theta)$, which are far from these walls. Consequently, the behavior of the system at the top of the barrier remains unchanged. The resulting ($\cos(\theta)$, Δd) histograms in the reactant well, along with the free energies from the umbrella sampling analysis, are presented in Figure 1, both with and without the wall on $\cos(\theta)$. The very slow stabilization of the degrees of freedom of the mass proves impractical at the *ab initio* level in the absence of walls.

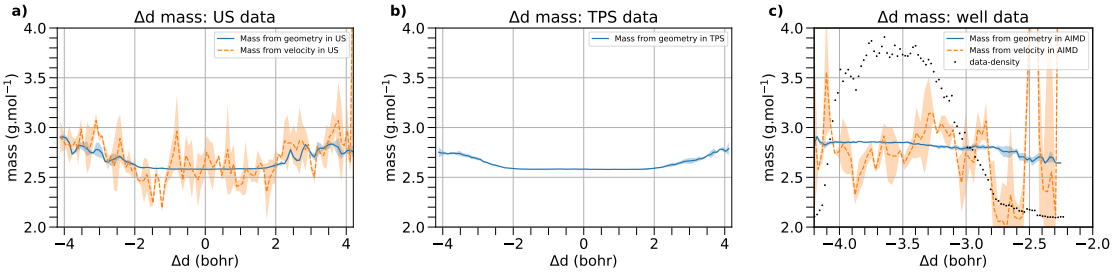


Figure 2: The effective mass estimated from all datasets with the wall on $\cos(\theta)$. For each system, the mass is estimated using Equation 4 with the values of $\cos(\theta)$ depicted in blue, or using the Maxwell-Boltzmann distribution shown in orange. Error bars are estimated using block averaging. For the US dataset, shown in sub-figure **a)**, the agreement between the two curves indicates the validity of Equation 13. The TPS dataset, with results presented in sub-figure **b)**, does not permit the use of the Maxwell-Boltzmann method as it does not follow the equilibrium velocity distribution. In the well ensemble, shown in sub-figure **c)**, the histogram of the trajectory is included and the agreement between the two curves is valid only in regions with a high density of points. It is observed that the velocity method requires significantly more time to converge, suggesting that the geometric method is more efficient for accurately quantifying the effective mass with feasible datasets.

The final effective mass, as a function of Δd, is presented in Figure 2 for all datasets, including the AIMD reactant well sampling. With the addition of the wall on $\cos(\theta)$, the effective mass of the system remains nearly constant with respect to Δd across all datasets.

Data-driven variables

In general, no suitable heuristic variable is known. To address this limitation, our previous study[4] on the same reaction introduced Path Collective Variables (PCV) derived from the TPS dataset.[35] This variable has been shown to accurately determine the barrier height.[4] However, since it is data-based, its performance is strongly dependent on the quality of the

dataset. In the present case, it loses the symmetry aspect of the reaction. The PCV pair is denoted as (s_{12}, z_{12}).

s_{12} and z_{12} exhibit complementary behaviors. s_{12} serves as the proposed reaction coordinate, varying from 1 to 12, where 2 corresponds to the reactant state and 11 to the product state. z_{12} is a distance coordinate that measures how far the system deviates from the reference set that defines s_{12}. Since this reference set is selected to represent the complete reaction mechanism, z_{12} remains close to 0 throughout the reaction, which may render it a less effective reaction coordinate.

This pair is trained using the element-wise coordination numbers of the carbon atom (C) and the two chlorine atoms (Cl^1, Cl^2), denoted as $c_A(\sigma)$, where A represents the selected atom and σ indicates the chosen element. The 12 reference frames used to define the PCV are available in the Supplementary Information (see **S-3**), comprising matrices of $c_A(\sigma)$ at various stages of the reaction, established from the shooting-from-the-top dataset using the protocol described in our previous paper.[4]

To compare with the heuristic variable Δd, we will examine the (s_{12}, z_{12}) pair, along with some of the coordination numbers used for their training. The final set of study variables is as follows:

- Δd
- s_{12}
- z_{12}
- d_1
- $c_C(K)$, the coordination number of the central carbon with the counter-ion K. This collective variable, which remains consistently close to 0, is likely uncorrelated with the transition.

- $c_{\mathrm{C}}(\mathrm{H})$, the coordination number of the central carbon with hydrogen. This collective variable exhibits strong memory and is also likely uncorrelated with the transition.

After analyzing these six variables, we conduct similar measurements on all coordination numbers used to generate s_{12} and z_{12} to calculate correlations with these parameters using a larger dataset.

Measured criteria

To compare these variables and assess their quality in relation to their dynamic behavior, we estimate the transmission coefficients (TC)[36,37] for each variable, using the TPS dataset. Additionally, we evaluate a quality criterion based on the committor, as derived from our previous study.[4] This criterion, referred to as the Standard Log-Likelihood Score (SLLS), is estimated through regression on committor values at the top of the barrier.[38] We present the variables derived from the models inferred from the reactant well ensemble:

- $t_{1/2}$ (fs): The numerical half-life of the memory kernel, estimated by integrating the memory kernel until it reaches half of its total integral.
- γ_1 (ps^{-1}): The friction coefficient obtained from linear regression of the friction.
- R_1^2 (-): The determination coefficient for γ_1.
- γ_2 (ps^{-1}): The friction coefficient derived from fitting the velocity auto-correlation.
- γ_3 (ps^{-1}): The friction coefficient obtained from the Kramer-Moyal equation.

Model Integration and Kinetic estimation

To extrapolate our results once a GLE or ULE model is established, we integrate it using a modified version of the freely available code by Jan Daldrop.[39] This code has been incorporated into a custom program for kinetics estimation, as detailed in the SI (see **S-5**). The

modification includes the addition of temperature as an external parameter. A study on the reproduction of dynamical data in the well to assess method stability is also described in the SI (see **S-6**). For underdamped integration, we use an internally developed code[40] employing the GJF integrator.[41]

In this paper, kinetics are estimated using two different methods: reactive flux[3,42] and T-boost.[43]

Reactive Flux is estimated from unbiased trajectories relaxing from the barrier top.[10] It accounts for the transmission coefficient at the top of the barrier as well as the barrier height itself, as described in Equation 14. In this study, reactive flux is applied directly to AIMD Transition Path Sampling data and also to data generated by GLE or underdamped integration that simulates transition path sampling. The reaction rate is estimated using the following equation:

$$\begin{aligned} \tau^{-1} &= \lim_{t_r \leq t \ll t_{fp}} k(t) \\ &= \lim_{t_r \leq t \ll t_{fp}} \langle v(0) h_B(t) \rangle \frac{e^{-\beta F^{\ddagger}}}{\int_A e^{-\beta F(x)} dx} \end{aligned} \tag{14}$$

where $h_B(t)$ denotes the indicator function for the product state. The prefactor of the exponential term is estimated using the mean value of the velocity at the top of the barrier, conditioned by $h_B(t)$. This velocity is considered the initial velocity of the trajectory since we perform shooting from the top. For values of t between the relaxation time from the top of the barrier, t_r (approximately 0.2 ps), and the MFPT, t_{fp} (approximately 1 minute), this mean value reaches a plateau where no new crossings occur, and it equals the prefactor. Since the MFPT for chemical reactions exceeds our computational capabilities at 300K during transition path sampling, t_{fp} remains significantly longer than our MD trajectories.

In the final expression, converging the estimate of the correlation function may require up to 10^5 unbiased trajectories initiated at the top of the barrier, depending on the system.

Particle velocities in molecular dynamics (MD) simulations are drawn from the Maxwell-Boltzmann distribution. Similarly, in Langevin dynamics simulations, the velocity of the CV is initialized from the Maxwell-Boltzmann distribution using the effective mass. A discussion on reactive flux in high-barrier phenomena can be found in Ghysbrecht's 2023 article.[1]

T-boost is a more direct method for estimating the reaction rate by using the first transition time of the system at elevated temperatures. This method determines the exponential and pre-exponential components of the rate via linear regression, as described by equations 15 and 16. It can be applied only with GLE or ULE integration, as the effect of temperature is constrained within these models.

$$\begin{aligned} \tau^{-1} &= Ae^{-\beta F^a} \\ &= Ae^{-\frac{F^a}{k_B T}} \end{aligned} \tag{15}$$

$$\implies \ln(\tau^{-1}) = \ln(A) - \frac{F^a}{k_B} T^{-1} \tag{16}$$

Since friction cannot be directly measured at the top of the barrier, we estimate the kinetics of the transition via the integration of a reconstructed GLE model.

$$\dot{v}(t) = f(x(t)) - \int_{-\infty}^{0} K(-\tau) v(t-\tau)\, d\tau + \xi(t) \tag{17}$$

where $f(x(t))$ is derived from the free energy landscape obtained through the WHAM applied on the US data, as shown in Equations 5 and 4. $K(\tau)$ denotes the memory kernel of the reactant well, which we truncated at 0.3 ps where the memory can be considered negligible according to our measurements. $\xi(t)$ represents colored Gaussian noise, determined from the memory kernel using the fluctuation-dissipation theorem.[44–46]

For underdamped integration, we use γ_1 as the friction coefficient, as it appears to be the best available approximation:

$$\dot{v}(t) = f(x(t)) - \gamma_1 v(t) + \xi(t) \tag{18}$$

where $\xi(t)$ is Gaussian white noise with an amplitude of $\sqrt{2\gamma_1 k_B T m^{-1}}$.

Results and discussion

GLE with stationary data

In order to assess the feasibility of modeling the behavior of a CV from a DFT calculation, we first applied our GLE inference algorithm to 50 ps of converged data from the reactant well (the first 50 ps were discarded to obtain the convergence). We selected a set of six CVs (see the Methods section) and plotted the corresponding memory kernels, as shown in Figure 3.

For each case, we estimated γ_1 (the linear correlation coefficient between $\int K_v\, dt$ and v), γ_2 by fitting the memory kernel using the SciPy toolkit in Python, and γ_3 using the Kramer-Moyal method. We compiled these values into a table along with the Standard Log-Likelihood Score (SLLS) and the transmission coefficient (TC), as shown in Table 1.

Table 1: Dynamics and quality features across a range of CVs with varying qualities. The SLLS quantifies how likely a variable is to be linearly transformed into the committor at the top of the barrier. The TC measures the recrossing effect during the transition. The subsequent parameters are derived from the GLE models and the data of the well. The friction coefficient and memory decay times are known to be inversely proportional to the quality of the variable to mimic the behavior of the committor. Error bars are estimated using block averages.

Var	SLLS	TC (%)	γ_1 (ps^{-1})	R_1^2 (-)	γ_2 (ps^{-1})	$t_{1/2}$ (fs)	γ_3 (ps^{-1})
Δd	17.2 ± 2.9	50.5 ± 2.2	76.3 ± 2.0	0.180 ± 0.001	20.3 ± 4.6	30.5 ± 3.4	15.4 ± 0.3
s_{12}	16.0 ± 2.5	41.0 ± 6.5	144.0 ± 3.2	0.094 ± 0.002	41.4 ± 0.1	35.6 ± 2.2	57.7 ± 1.0
z_{12}	5.81 ± 5.4	7.5 ± 1.6	160.0 ± 3.1	0.099 ± 0.013	60.5 ± 15.6	31.9 ± 2.4	70.4 ± 3.2
d_1	14.5± 2.7	50.8 ± 3.3	51.8 ± 1.2	0.43 ± 0.02	22.8 ± 12.9	76.7 ± 16.7	8.2 ± 0.2
c_C(K)	1.6 ± 1.6	0.6 ± 3.8	89.7 ± 3.3	0.175 ± 0.007	11.1 ± 9.1	71.1 ± 3.4	19.0 ± 1.1
c_C(H)	1.0 ± 0.2	4.5 ± 5.1	6.5 ± 1.3	0.001 ± 0.001	14.2 ± 0.1	348.6± 69.9	39.7 ± 5.7

The linear correlation coefficients derived from this study are presented in the following

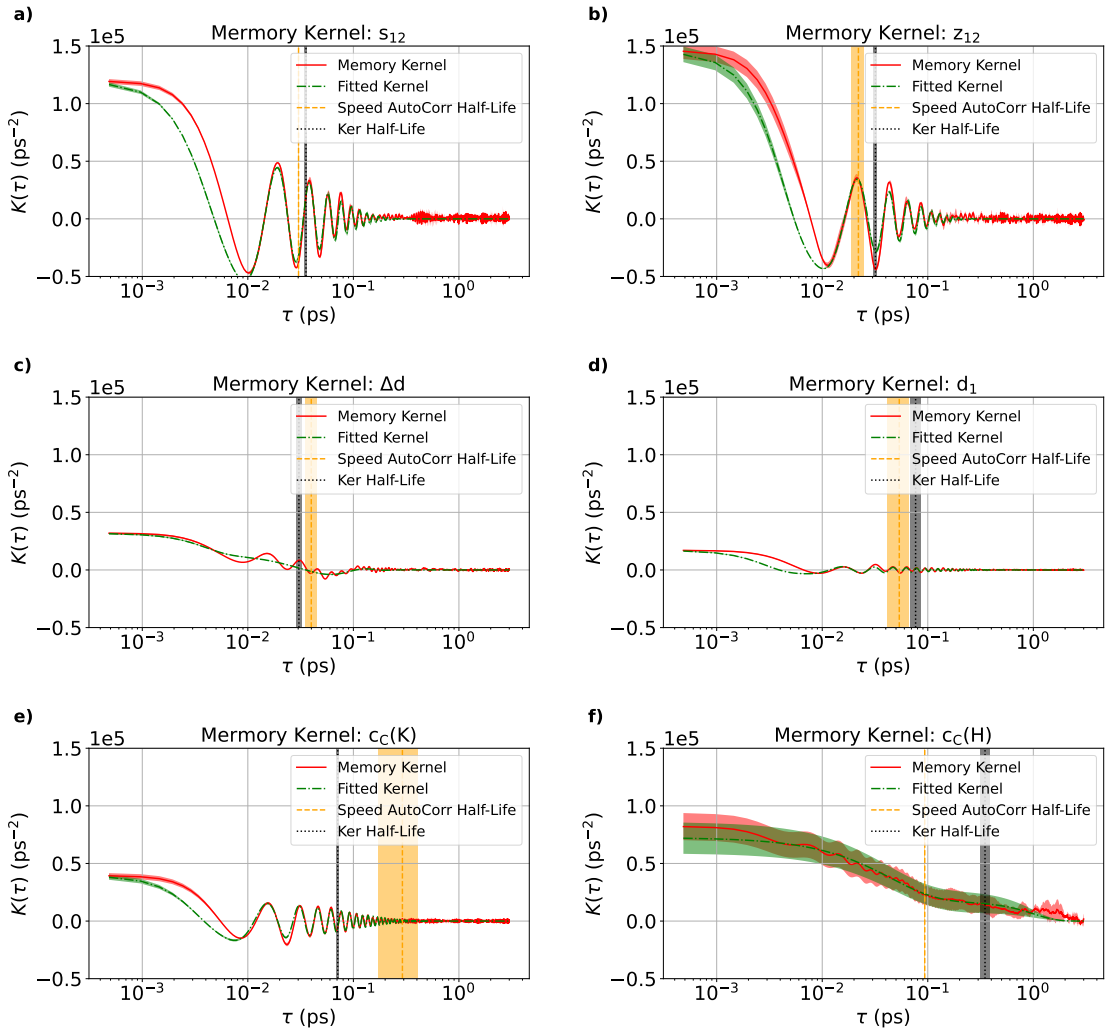


Figure 3: Memory kernels for the S_N2 reaction in the reactant well using s_{12}, z_{12}, Δd, d_1, c_C-K, and c_C-H. The fits are represented by dashed green lines. The memory kernel half-life is shown in gray, and the auto-correlation of the velocity half-life is shown in orange. Most cases show that the memory kernel half-time is larger than or comparable to the auto-correlation half-time, justifying that the friction coefficient cannot be estimated by a straightforward integration of the memory. Error bars are estimated using block average.

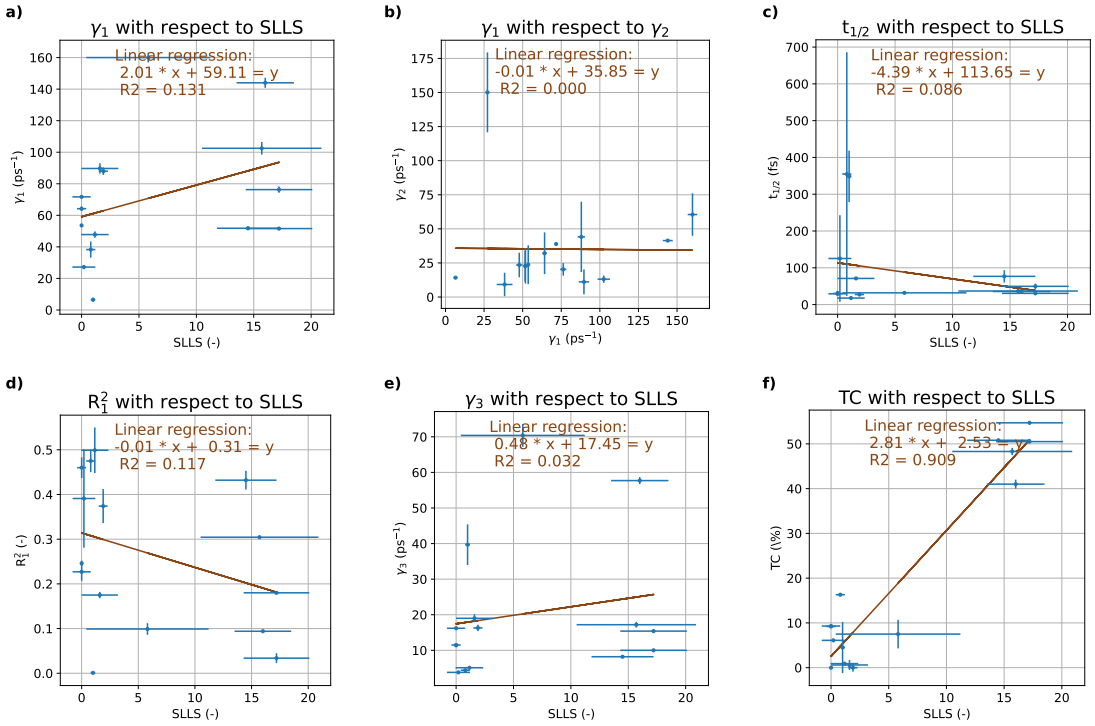


Figure 4: Graphs of the data from Table 1 and additional variables detailed in the SI (see **S-4**) with linear regression analysis. These graphs illustrate that the SLLS is decorrelated from the friction coefficients (γ_1, γ_2, and γ_3), the half-life ($t_{1/2}$) of the memory kernel, and the determination coefficient (R_1^2) of the friction coefficient γ_1. However, SLLS is correlated with the TC.

matrix:

$$\begin{bmatrix} & SLLS & \gamma_1 & R_1^2 & \gamma_2 & t_{1/2} & \gamma_3 & TC \\ SLLS & 1.00 & 0.36 & -0.34 & -0.22 & -0.29 & 0.18 & 0.95 \\ \gamma_1 & & 1.00 & -0.28 & -0.01 & -0.55 & 0.70 & 0.17 \\ R_1^2 & & & 1.00 & 0.13 & -0.04 & -0.64 & -0.21 \\ \gamma_2 & & & & 1.00 & -0.11 & 0.03 & -0.24 \\ t_{1/2} & & & & & 1.00 & -0.03 & -0.16 \\ \gamma_3 & & & & & & 1.00 & -0.02 \\ TC & & & & & & & 1.00 \end{bmatrix}$$

The correlation matrix indicates that the SLLS is not correlated with any of the friction coefficients. Similarly, there is no significant correlation with the decay time of memory the kernel ($t_{1/2}$) in the well. This suggests that friction and memory time in the well may not be reliable criteria for optimizing variables to achieve a good regression of the committor. However, as expected, SLLS does show a correlation with the TC, reflecting their mutual dependence on the behavior of the CV at the top of the barrier.

To improve the evaluation, another criterion is needed that considers not only the behavior of the variable at the top of the barrier but also its behavior throughout the entire transition.

Kinetic energies of the CVs

To gain deeper insights into the behavior of our CV during the reaction, we can perform a complementary analysis on the shooting-from-top dataset. In the literature, an optimal CV is often defined as one that effectively captures and dissipates the kinetic energy released during barrier crossing. This phenomenon can be interpreted as a "warming-up" effect of the variable. The term $\frac{1}{2}m_{\text{eff}}\langle v^2\rangle$ estimates the kinetic energy for each variable. At equilibrium, this quantity remains constant according to the Maxwell-Boltzmann distribution. However, during barrier descent, this value may change if the variable significantly influences the

reaction mechanism.

If N shots are performed from the top, and assuming that the masses of the studied CVs remain constant near the top of the barrier region, we can plot the time evolution of this "warming-up" effect for each CV in our set. This is achieved by calculating the ensemble mean value of all shots and rescaling it by the mean value at $t = 0$ to account for initial mass effects. The calculation is performed as follows:

$$Ekin_{cv}(t) = \frac{\langle v_{cv}^2(t) \rangle}{\langle v_{cv}^2(0) \rangle} \tag{19}$$

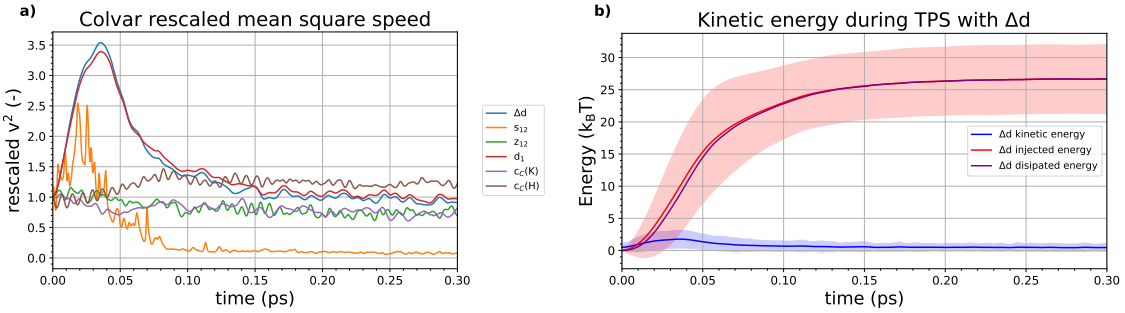


Figure 5: Kinetic study of the variables during shooting from the top. **a)** Rescaled $\langle v^2 \rangle(t)$ for each CV relative to the initial distribution, as described in Equation 19. It can be observed that the heuristic variable Δd absorbs the most heat during the descent. The pronounced oscillations in the s_{12} curve likely indicate significant fluctuations in the effective mass of this data-driven variable during the descent, which is undesirable for dynamic studies. The high time asymptotes of the curves that are not equal to one suggest variations in the effective mass between the top of the barrier and the wells. **b)** Estimation of the total reduced energy transferred to the Δd variable using the free energy curve from US data and the effective mass at the top of the barrier. It is evident that, despite the velocity distribution having a non-zero mean value during the descent, this variable efficiently dissipates most of the energy injected by the potential.

In Figure 5, we reaffirm our previous findings that optimizing only the memory aspect of a variable (that is only available in the well) is insufficient to ensure its quality. A significant aspect of this study is its analysis of the behavior of the variables throughout the entire reaction, rather than focusing solely on the barrier top (e.g., SLLS, reactive flux) or the well ensemble (e.g., memory kernels, R_1^2, γ_1). A notable distinction emerges between the

variables s_{12} and Δd: the s_{12} curve exhibits pronounced oscillations, while the Δd curve remains smoother. These oscillations likely reflect variations in the effective mass of s_{12} as the system transitions between different reference frames. This observation suggests that data-driven variables like s_{12} may require refinement to achieve smoother behavior in their associated effective mass curves, which is crucial for their use in future dynamic studies. This underscores the need for ongoing improvements in the development of data-driven collective variables (CVs) for chemical reaction kinetics.

Given its favorable properties observed thus far, we proceed to use Δd for rate estimation.

Kinetic estimation

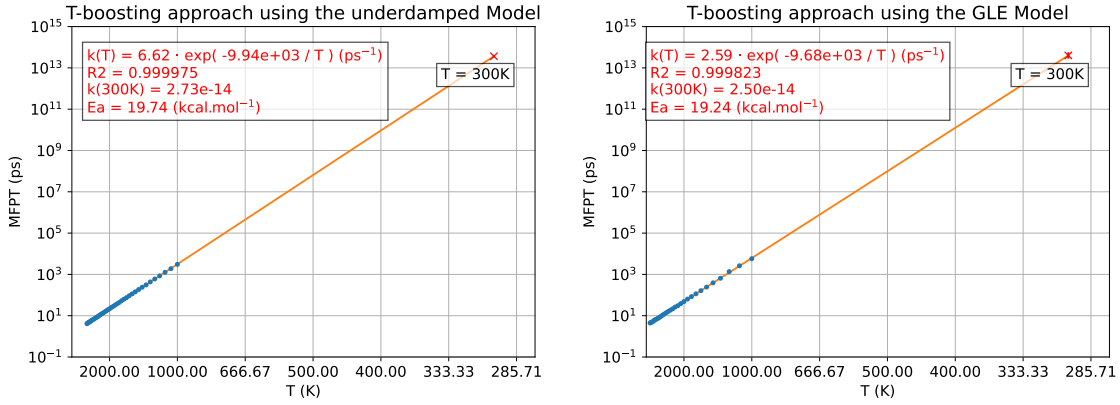


Figure 6: Kinetics estimation using the T-boost method: **a)** Results from the underdamped model and **b)** Results from the generalized model. The two models exhibit good agreement regarding the final kinetic estimation. The statistical deviations of the data points are too small to be visible and are accounted for in the linear regression. Error bars for the 300 K value of the MFPT are estimated based on the standard deviation of the coefficients a and b in the model.

Table 2 presents the results of our rate estimation for the reaction using our methods. We employed the reconstructed model described in the Methods section, incorporating a free energy profile derived from Umbrella Sampling (US) analysis and a constant memory kernel based on data from the well. Using these parameters, we estimated the rate using both the reactive flux and T-boost methods. Our estimated rate is $2.5 \cdot 10^{-2}$ s^{-1}, compared to the

Table 2: Estimation of the transition rate using all our protocols and models. The error bars are determined using block averaging for reactive flux and standard deviation for T-boost. All measurements agree on the order of magnitude of the rate: approximately $1 \cdot 10^{-2}$ s^{-1}.
* with initialization of the memory set to zero

method	rate	transmission
GLE T-boost	(2.5 ± 1.1)e-2 (s^{-1})	NA
underdamped T-boost	(2.7 ± 0.4)e-2 (s^{-1})	NA
GLE Reactivflux*	(1.5 ± 0.1)e-2 (s^{-1})	0.20 ± 0.01
underdamped Reactivflux	(2.2 ± 0.1)e-2 (s^{-1})	0.29 ± 0.01
DFT Reactivflux	(3.7 ± 1)e-2 (s^{-1})	0.49 ± 0.2
Experimental	$\approx$5e-7 (s^{-1})[47]	NA

experimental estimate of $2.5 \cdot 10^{-7}$ s^{-1}. Given that our free-energy barrier is underestimated (21.4 kcal/mol as opposed to 26 kcal/mol reported in the literature[14]), the obtained rate is reasonable. A simple estimation using $\frac{k_B T}{h} e^{-\beta \Delta F}$ yields approximately 10^{-3} s^{-1} with our measured barrier height, while using $\Delta F = 26$ kcal/mol gives approximately 10^{-6} s^{-1}. This indicates an overall agreement between the reactive flux and T-boost results in terms of orders of magnitude.

In the case of the S_N2 reaction and Δd, the generalization of the friction from the well to the top of the barrier warrants discussion. We conducted a study suggesting that friction and the memory kernel are lower at the top of the barrier; details are provided in the SI (see **S-9**). However, accurately inferring the true memory kernel in these regions remains challenging. To achieve static conditions, biases would need to be introduced in these regions, and we currently lack a method to precisely calculate their impact on the memory kernel. Moreover, we do not have a sufficiently high-quality *ab initio* dataset suitable for training a position-dependent memory kernel,[9] as might be achievable with some classical molecular dynamics approaches.[48,49]

Conclusion

For dynamic behavior inference, the requirements for a "good" collective variable (CV) are more stringent than those for static barrier height determination, as in Umbrella Sampling

(US). It is insufficient for a CV merely to describe the intermediate stages of a reaction and determine the true height of the barrier. Additional properties are essential. One such property identified in this work is a smooth effective mass curve during the transition. However, constructing such variables is considerably more challenging with data-driven CVs, which encompass many degrees of freedom in the bath and tend to aggregate numerous contributions from the solvent degrees of freedom in their effective mass.

Using the S_N2 standard reaction and the MCCI system, we conducted a comprehensive evaluation of different variable types, highlighting two significant observations:

1. The memory characteristic time and the friction coefficient (estimated through regression of the friction against the velocity) in the well ensemble are not correlated with the SLLS quality criteria or the transmission coefficient at the top of the barrier. This suggests that optimizing the memory kernel based on well data alone should not be the sole criterion for identifying a good CV.

2. Data-driven variables, such as s_{12}, which demonstrate a good SLLS score and a favorable barrier height in Umbrella Sampling, can still exhibit substantial variations in their effective mass. These variations are challenging to stabilize due to the multiple degrees of freedom involved, rendering such variables unsuitable for dynamic inference. Thus, a good barrier height alone is not a sufficient criterion for optimization in dynamic studies.

For Δd, even with a single "solvent" degree of freedom (θ) in the mass estimation, it was necessary to introduce a wall to achieve mass convergence. This addition allowed us to treat the system's mass as quasi-constant, thereby reducing errors when generalizing the friction and noise measurements from the well to the rest of the reaction. In more complex systems where no heuristic reaction coordinate is known, determining a data-driven reaction coordinate with a deterministic or quickly converging effective mass could be intricate. In such cases, the use of external walls, as implemented in our study, helps identify all relevant

degrees of freedom of the mass. However, this addition can significantly alter the system's behavior, potentially rendering kinetic results unreliable.

By combining Umbrella Sampling for free energy and well equilibrium data for friction estimation, we developed GLE and ULE models that produced reasonable rate estimates based on the measured barrier. Scaling the barrier height with experimental data brought us closer to experimental results.

References

References

(1) Ghysbrecht, S.; Donati, L.; Keller, B. G. Accuracy of reaction coordinate based rate theories for modelling chemical reactions: insights from the thermal isomerization in retinal. 2023; `http://arxiv.org/abs/2312.12948`, arXiv:2312.12948 [physics].

(2) Hijón, C.; Español, P.; Vanden-Eijnden, E.; Delgado-Buscalioni, R. Mori–Zwanzig formalism as a practical computational tool. *Faraday Discussions* **2010**, *144*, 301–322, Publisher: Royal Society of Chemistry.

(3) Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **1940**, *7*, 284–304.

(4) Magrino, T.; Huet, L.; Saitta, A. M.; Pietrucci, F. Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry. *The Journal of Physical Chemistry A* **2022**, *126*, 8887–8900, Publisher: American Chemical Society.

(5) Ensing, B.; Meijer, E. J.; Blöchl, P. E.; Baerends, E. J. Solvation Effects on the SN2 Reaction between CH3Cl and Cl- in Water. *The Journal of Physical Chemistry A* **2001**, *105*, 3300–3310, Number: 13 Publisher: American Chemical Society.

(6) Berne, B. J.; Borkovec, M.; Straub, J. E. Classical and modern methods in reaction rate theory. *The Journal of Physical Chemistry* **1988**, *92*, 3711–3725.

(7) Benjamin, I.; Lee, L. L.; Li, Y. S.; Liu, A.; Wilson, K. R. Generalized Langevin model for molecular dynamics of an activated reaction in solution. *Chemical Physics* **1991**, *152*, 1–12.

(8) Berkowitz, M.; Morgan, J. D.; McCammon, J. A. Generalized Langevin dynamics simulations with arbitrary time-dependent memory kernels. *The Journal of Chemical Physics* **1983**, *78*, 3256–3261.

(9) Vroylandt, H.; Monmarché, P. Position-dependent memory kernel in generalized Langevin equations: Theory and numerical estimation. *The Journal of Chemical Physics* **2022**, *156*, 244105.

(10) Palacio-Rodriguez, K.; Pietrucci, F. Free Energy Landscapes, Diffusion Coefficients, and Kinetic Rates from Transition Paths. *Journal of Chemical Theory and Computation* **2022**, *18*, 4639–4648, Publisher: American Chemical Society.

(11) Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annual Review of Physical Chemistry* **2016**, *67*, 669–690, _eprint: https://doi.org/10.1146/annurev-physchem-040215-112215.

(12) Elliott, S.; Rowland, F. S. Nucleophilic substitution rates and solubilities for methyl halides in seawater. *Geophysical research letters* **1993**, *20*, 1043–1046.

(13) Mathis, J. R.; Bianco, R.; Hynes, J. T. On the activation free energy of the Cl- + CH3Cl SN2 reaction in solution. *Journal of Molecular Liquids* **1994**, *61*, 81–101.

(14) Albery, W. J.; Kreevoy, M. M. In *Advances in Physical Organic Chemistry*; Gold, V., Bethell, D., Eds.; Academic Press, 1978; Vol. 16; pp 87–157.

(15) McLennan, D. J. Semi-empirical calculation of rates of SN2 Finkelstein reactions in solution by a quasi-thermodynamic cycle. *Australian Journal of Chemistry* **1978**, *31*, 1897–1909, Publisher: CSIRO PUBLISHING.

(16) Grote, R. F.; Hynes, J. T. The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *The Journal of Chemical Physics* **1980**, *73*, 2715–2732.

(17) Tirado-Rives, J.; Jorgensen, W. L. QM/MM Calculations for the Cl $^-$ + CH $_3$ Cl S $_N$ 2 Reaction in Water Using CM5 Charges and Density Functional Theory. *The Journal of Physical Chemistry A* **2019**, *123*, 5713–5717, Number: 27.

(18) Hutter, J.; Curioni, A. Car–Parrinello Molecular Dynamics on Massively Parallel Computers. *ChemPhysChem* **2005**, *6*, 1788–1793, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.200500059.

(19) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996**, *77*, 3865–3868, Publisher: American Physical Society.

(20) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006**, *27*, 1787–1799, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20495.

(21) Troullier, N.; Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Physical Review B* **1991**, *43*, 1993–2006, Publisher: American Physical Society.

(22) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **1985**, *31*, 1695–1697, Publisher: American Physical Society.

(23) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **1984**, *81*, 511–519.

(24) Martyna, G. J.; Klein, M. L.; Tuckerman, M. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *The Journal of Chemical Physics* **1992**, *97*, 2635–2643.

(25) Jung, B.; Jung, G. Dynamic Coarse-Graining of Linear and Non-Linear Systems: Mori-Zwanzig Formalism and Beyond. 2023; `http://arxiv.org/abs/2307.08143`, arXiv:2307.08143 [cond-mat].

(26) Kästner, J. Umbrella sampling. *WIREs Computational Molecular Science* **2011**, *1*, 932–942, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.66.

(27) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.11. `http://membrane.urmc.rochester.edu/?page_id=126`.

(28) Vroylandt, H. On the derivation of the generalized Langevin equation and the fluctuation-dissipation theorem. *Europhysics Letters* **2022**, *140*, 62003, Publisher: EDP Sciences, IOP Publishing and Società Italiana di Fisica.

(29) Vroylandt, H.; Goudenège, L.; Monmarché, P.; Pietrucci, F.; Rotenberg, B. Likelihood-based non-Markovian models from molecular dynamics. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2117586119, arXiv:2110.04246 [cond-mat, physics:physics].

(30) Mouaffac, L.; Palacio-Rodriguez, K.; Pietrucci, F. Optimal Reaction Coordinates and Kinetic Rates from the Projected Dynamics of Transition Paths. *Journal of Chemical Theory and Computation* **2023**, *19*, 5701–5711, Publisher: American Chemical Society.

(31) Chorin, A. J.; Hald, O. H.; Kupferman, R. Optimal prediction with memory. *Physica D: Nonlinear Phenomena* **2002**, *166*, 239–257.

(32) HadrienNU/VolterraBasis: Python 3 tool suite for the computation of posi-

tion dependent memory kernels from time series. https://github.com/HadrienNU/VolterraBasis.

(33) Linz, P. Numerical methods for Volterra integral equations of the first kind. *The Computer Journal* **1969**, *12*, 393–397.

(34) Girardier, D. D.; Vroylandt, H.; Bonella, S.; Pietrucci, F. Inferring free-energy barriers and kinetic rates from molecular dynamics via underdamped Langevin models. *The Journal of Chemical Physics* **2023**, *159*, 164111.

(35) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *The Journal of Chemical Physics* **2007**, *126*, 054103.

(36) Devergne, T. Machine learning methods for computational studies in origins of life, Thèse de Doctorat de Physique.

(37) Palacio-Rodriguez, K.; Vroylandt, H.; Stelzl, L. S.; Pietrucci, F.; Hummer, G.; Cossio, P. Transition Rates and Efficiency of Collective Variables from Time-Dependent Biased Simulations. *The Journal of Physical Chemistry Letters* **2022**, *13*, 7490–7496, Publisher: American Chemical Society.

(38) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *The Journal of Chemical Physics* **2006**, *125*, 054108.

(39) Daldrop, J. jandaldrop/bgle. 2024; https://github.com/jandaldrop/bgle, original-date: 2018-07-10T16:47:44Z.

(40) Hadrien HadrienNU/LangevinIntegrators.jl. 2023; https://github.com/HadrienNU/LangevinIntegrators.jl, original-date: 2022-08-17T20:28:56Z.

(41) Grønbech-Jensen, N.; Farago, O. A simple and effective Verlet-type algorithm for simulating Langevin dynamics. *Molecular Physics* **2013**, *111*, 983–991, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00268976.2012.760055.

(42) Chandler, D. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *The Journal of Chemical Physics* **2008**, *68*, 2959–2970.

(43) Wolf, S.; Lickert, B.; Bray, S.; Stock, G. Multisecond ligand dissociation dynamics from atomistic simulations. *Nature Communications* **2020**, *11*, 2918, Number: 1 Publisher: Nature Publishing Group.

(44) Maes, C.; Safaverdi, S.; Visco, P.; van Wijland, F. Fluctuation-response relations for nonequilibrium diffusions with memory. *Physical Review E* **2013**, *87*, 022125, Publisher: American Physical Society.

(45) Lü, K.; Bao, J.-D. Numerical simulation of generalized Langevin equation with arbitrary correlated noise. *Physical Review E* **2005**, *72*, 067701, Publisher: American Physical Society.

(46) Schmidt, J.; Meistrenko, A.; van Hees, H.; Xu, Z.; Greiner, C. Simulation of stationary Gaussian noise with regard to the Langevin equation with memory effect. *Physical Review E* **2015**, *91*, 032125, Publisher: American Physical Society.

(47) Baesman, S. M.; Miller, L. G. Laboratory Determination of the Carbon Kinetic Isotope Effects (KIEs) for Reactions of Methyl Halides with Various Nucleophiles in Solution. *Journal of Atmospheric Chemistry* **2005**, *52*, 203–219.

(48) Díaz Leines, G.; Rogal, J. Maximum Likelihood Analysis of Reaction Coordinates during Solidification in Ni. *The Journal of Physical Chemistry B* **2018**, *122*, 10934–10942, Publisher: American Chemical Society.

(49) Rogal, J.; Lechner, W.; Juraszek, J.; Ensing, B.; Bolhuis, P. G. The reweighted path ensemble. *The Journal of Chemical Physics* **2010**, *133*, 174109.

6.2 Conclusion in the context of the Thesis

This final article concludes the results presented in this work. We succeeded in producing a study in which these tools are discussed, confronted, and used for the inference of the kinetics of a chemical reaction in solution. From an external perspective, this article serves as a valuable entry point to encourage discussions between theoretical chemists and stochastic physicists, aiming to develop these promising tools for accurate and agnostic estimation of the kinetics for chemical reactions in solution, or within a broader framework for high-barrier transformations in condensed matter physics.

Another key feature highlighted in this article is that the main source of error in the rate estimation remains due to an inaccurate determination of the barrier height, which is based on the method we use to calculate the forces in AIMD. Future studies in this area should also focus on developing new methods to enhance the level of theory in force calculation, thereby achieving realistic values for kinetics.

Part III

Conclusion

Conclusion

During this thesis, we followed the path of an *ab initio* framework for the study of chemical reactions in solution, both for thermodynamics and kinetics. For the thermodynamic aspect, we demonstrate that the use of a collective variable driven by data, combined with a neural network potential, allows efficient determination of the thermodynamic constants of a chemical pathway completely discovered *in silico*. The robustness of the final workflow was validated by performing two coherent estimations of the same free energy difference for the same synthesis via two different pathways. Both syntheses were complex, involving 6 to 7 elementary steps, each treated individually.

The results of this study provide new insights into the broad field of prebiotic chemistry by detailing new intermediates of an unexplored pathway for amino acid synthesis. In addition, it offers valuable and reliable data on the thermodynamics of these processes. These results can be compared with experimental measurements, ensuring the refutability of the findings, in line with good scientific practice.

With regard to kinetics, our study demonstrates that GLE is a promising tool for establishing a reliable stochastic dynamics of a reaction in solution. It is feasible to derive kinetics from this model using various methods whose results are in good agreement. However, the application of these tools in a perfectly agnostic framework, which accounts for the variations of the friction along the reaction and the effective mass, remains an open challenge.

Perspectives

Concerning the data-driven path collective variables, the use of a low-angle criterion to optimize the path is a feature that could be discussed. This choice often leads to selecting data points at the borders of the "reactive cone," especially if the MFEP presents large angles in the chemical space. It could be potentially replaced by other criteria that would still encourage low distances between the references that form the path to efficiently target the MFEP. One possible idea to develop in the future could be a "property-oriented" PCV, optimized with two criteria: homogeneous distances, always a mandatory criterion in the field, and the lowest possible value of z_{12} across the dataset. Optimizing these two criteria would take advantage of an important property of PVC, where z_{12} measures the distance to the path. Reducing z_{12} for the entire dataset would, by definition, lead to a path close to the center of the "reactive cone".

To test this hypothesis, a new protocol should be developed that uses the same optimization process as the one used in this thesis, based on a Monte Carlo algorithm, or maybe a more recent version of optimization based on machine learning. This new estimation of a Pathcv could be compared with the actual one using the same dataset as that used in the GLE article. This could also be tested first with synthetic data in order to have a first insight of the challenge and feasibility of this new type of PCVs.

Concerning GLE, I see two different solutions for the issues we have to deal with. For the variation of friction with respect to position, we have no choices but to change the dataset we work with. The problem with the Umbrella Sampling dataset is that it comprises biased trajectories with unknown effects on memory. The issue with the Transition Path Sampling dataset is that it consists of unequilibrated trajectories from the top. These trajectories can not be used to infer memory because they are not ergodic. A rescaling factor of the weights of each trajectories in the dataset could allow for use to recover the ergodic distribution even for trajectories that cross the barrier. This can be obtained by using a different kind of dataset, such as the reweighted path ensemble[66]. However, the quantity of data that compose these datasets would be significantly larger than the ones we have used so far, which may be challenging in a fully AIMD framework. We can try to obtain it using MLMD.

If this kind of reweighted dataset can be generated for a chemical reaction, then the different weights of each trajectory can be used inside the estimation of the GLE parameters that are the mean forces and the memory kernel, especially to weight all the

estimated averages. By doing that, we can recover a full GLE model of the reaction. If the preceding assertion are revealed to be possible then it could open the door to many remaining questions. Among them, the importance of the usage of a position-dependent memory kernel and on how it could be implemented and integrated. Obtaining a reliable model will also help us demonstrate the utility of these tools by calculating the kinetics of a chemical reaction from an unbalanced dataset and comparing these results with those from smaller models such as Underdamped Langevin Equation (ULE).

An other issue has to be addressed in parallel to the establishment of a complete gle model of a chemical reaction, it is the mass variation during the reaction, especially the one of data-driven PCVs.

To avoid pics in the mass of PCV we could return to the fundamentals and revisit the continuous formulation of PCV[64]. This approach could ensure smooth behavior without jumping in mass between reference points. In order to do that, we can use a continuous model like the Lagrangian polynôme to make the path of matrices continuous. This could be tested with some coding efforts on the shooting form the top dataset that has been used to discover the pics. It could also be projected on an easily integrable basis of function in order to permit "on the fly" estimation during molecular dynamics.

Furthermore, to project the remaining variation to a constant mass we could modify the index function (the one that initially counts the references)[64] to rescale the mass of s_{12} and maintain its consistency along the path. This index function could be renamed the reference density function, as it deals with the quantity of progression in the reference path when s evolves. Making this abrupte local will diminish the mass of the corresponding region of s_{12}. The establishment of the continuous PCV and reference density function would also benefit from using the reweighted path set as a database.

These perspectives call for larger collaborations between the fields of molecular dynamics and stochastic dynamics to develop accurate tools for inferring kinetics in condensed phases. In particular, the issues encountered in the last study pave the way to finally generate a database CV that could be a transferable and reliable tool for kinetic inference of chemical reactions in solution.

Bibliography

[1] T. Limpanuparb, D. Sathainthammanee, P. Pakwilaikiat, C. Kaewpichit, W. Yimkosol, and A. Suwannakhan, "Reinterpreting Popular Demonstrations for Use in a Laboratory Safety Session That Engages Students in Observation, Prediction, Record Keeping, and Problem Solving," *Journal of Chemical Education*, vol. 98, pp. 191–197, Jan. 2021. Publisher: American Chemical Society.

[2] J. Perrin, *Atoms*. London : Constable, 1916.

[3] P. W. Anderson, "More Is Different," *Science*, vol. 177, pp. 393–396, Aug. 1972.

[4] F. D. Drake, *Is anyone out there? : the scientific search for extraterrestrial intelligence*. New York, N.Y. : Delta Book/Dell Pub., 1994.

[5] A. Brack, "Liquid water and the origin of life," *Origins of life and evolution of the biosphere*, vol. 23, pp. 3–10, Feb. 1993.

[6] T. Devergne, *Machine learning methods for computational studies in origins of life*. phdthesis, Sorbonne Université, Sept. 2023.

[7] H. J. Cleaves, "Prebiotic Chemistry: Geochemical Context and Reaction Screening," *Life*, vol. 3, pp. 331–345, June 2013.

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

[9] S. L. Miller, "Production of Some Organic Compounds under Possible Primitive Earth Conditions [1]," *Journal of the American Chemical Society*, vol. 77, pp. 2351–2361, May 1955.

[10] A. M. Saitta and F. Saija, "Miller experiments in atomistic computer simulations," *Proceedings of the National Academy of Sciences*, vol. 111, no. 38, pp. 13768–13773, 2014. Publisher: National Academy of Sciences _eprint: http://www.pnas.org/content/111/38/13768.full.pdf.

[11] R. Krishnamurthy, S. Pulletikurti, M. Yadav, and G. Springsteen, "Prebiotic Synthesis of alpha-Amino Acids and Orotate from alpha-Ketoacids Potentiates Transition to Extant Metabolic Pathways," preprint, In Review, Oct. 2021.

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

[13] 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.

[14] T. Magrino, F. Pietrucci, and A. M. Saitta, "Step by Step Strecker Amino Acid Synthesis from Ab Initio Prebiotic Chemistry," *The Journal of Physical Chemistry Letters*, vol. 12, pp. 2630–2637, Mar. 2021. Publisher: American Chemical Society.

[15] J. W. Gibbs, *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Cambridge Library Collection - Mathematics, Cambridge: Cambridge University Press, 2010.

[16] M. J. Klein, "The Physics of J. Willard Gibbs in his Time," *Physics Today*, vol. 43, pp. 40–48, Sept. 1990.

[17] A. D. Kirwan, "Intrinsic photon entropy? The darkside of light," *International Journal of Engineering Science*, vol. 42, pp. 725–734, Apr. 2004.

[18] G. K. Batchelor, *An Introduction to Fluid Dynamics*. Cambridge Mathematical Library, Cambridge: Cambridge University Press, 2000.

[19] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*. Cambridge: Cambridge University Press, 2 ed., 2020.

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

[21] G. J. Martyna, M. L. Klein, and M. Tuckerman, "Nosé–Hoover chains: The canonical ensemble via continuous dynamics," *The Journal of Chemical Physics*, vol. 97, pp. 2635–2643, Aug. 1992.

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

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

[24] L. Boltzmann, "On Statistical Mechanics," in *Theoretical Physics and Philosophical Problems: Selected Writings* (L. Boltzmann and B. McGuinness, eds.), pp. 159–172, Dordrecht: Springer Netherlands, 1974.

[25] 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.

[26] J. C. Tully, "Perspective on "Zur Quantentheorie der Molekeln"," *Theoretical Chemistry Accounts*, vol. 103, pp. 173–176, Feb. 2000.

[27] A. Sarai and D. DeVault, "[5] Proton tunneling," in *Methods in Enzymology*, vol. 127 of *Biomembranes Part O: Protons and Water: Structure and Translocation*, pp. 79–91, Academic Press, Jan. 1986.

[28] S. Xu, S. Deng, L. Ma, Q. Shi, M. Ge, and X. Zhang, "The proton-coupled proton transfer mechanism, H2O catalysis, and hydrogen tunneling effects in the reaction of HNCH2 with HCOOH in the interstellar medium," *International Journal of Quantum Chemistry*, vol. 110, no. 14, pp. 2671–2682, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.22386.

[29] X. Meng, J. Guo, J. Peng, J. Chen, Z. Wang, J.-R. Shi, X.-Z. Li, E.-G. Wang, and Y. Jiang, "Direct visualization of concerted proton tunnelling in a water nanocluster," *Nature Physics*, vol. 11, pp. 235–239, Mar. 2015. Publisher: Nature Publishing Group.

[30] K. Umesaki and K. Odai, "Tunneling Effect in Proton Transfer: Transfer Matrix Approach," *The Journal of Physical Chemistry A*, vol. 127, pp. 1046–1052, Feb. 2023. Publisher: American Chemical Society.

[31] Collaboration: Scientific Group Thermodata Europe (SGTE), "Thermodynamic Properties of Compounds, CuS to ErF3," in *Pure Substances. Part 3 _ Compounds from CoCl3_g to Ge3N4* (Lehrstuhl für Theoretische Hüttenkunde and Rheinisch-Westfälische Technische Hochschule Aachen, eds.), vol. 19 A3, pp. 175–200, Berlin/Heidelberg: Springer-Verlag, 2000. Series Title: Landolt-Börnstein - Group IV Physical Chemistry.

[32] R. P. Feynman, "Forces in Molecules," *Physical Review*, vol. 56, pp. 340–343, Aug. 1939. Publisher: American Physical Society.

[33] 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.

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

[35] P. Lykos and G. W. Pratt, "Discussion on The Hartree-Fock Approximation," *Reviews of Modern Physics*, vol. 35, pp. 496–501, July 1963. Publisher: American Physical Society.

[36] L. H. Thomas, "The calculation of atomic fields," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 23, pp. 542–548, Jan. 1927.

[37] E. Fermi, "Statistical method to determine some properties of atoms," *Rend. Accad. Naz. Lincei*, vol. 6, no. 5, pp. 602–607, 1927.

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

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

[40] J. P. Perdew, "Climbing the ladder of density functional approximations," *MRS Bulletin*, vol. 38, pp. 743–750, Sept. 2013.

[41] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis," *Canadian Journal of Physics*, vol. 58, pp. 1200–1211, Aug. 1980. Publisher: NRC Research Press.

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

[43] J. P. Perdew and A. Zunger, "Self-interaction correction to density-functional approximations for many-electron systems," *Physical Review B*, vol. 23, pp. 5048–5079, May 1981. Publisher: American Physical Society.

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

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

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

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

[48] K. E. Riley, B. T. Op't Holt, and K. M. Merz, "Critical Assessment of the Performance of Density Functional Methods for Several Atomic and Molecular Properties," *Journal of Chemical Theory and Computation*, vol. 3, pp. 407–433, Mar. 2007. Publisher: American Chemical Society.

[49] C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The PBE0 model," *The Journal of Chemical Physics*, vol. 110, pp. 6158–6170, Apr. 1999.

[50] J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations," *The Journal of Chemical Physics*, vol. 105, pp. 9982–9985, Dec. 1996.

[51] M. Ernzerhof and G. E. Scuseria, "Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional," *The Journal of Chemical Physics*, vol. 110, pp. 5029–5036, Mar. 1999.

[52] J. Hutter and A. Curioni, "Car–Parrinello Molecular Dynamics on Massively Parallel Computers," *ChemPhysChem*, vol. 6, no. 9, pp. 1788–1793, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.200500059.

[53] T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, and J. Hutter, "CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations," *The Journal of Chemical Physics*, vol. 152, p. 194103, May 2020.

[54] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, "Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach," *Computer Physics Communications*, vol. 167, pp. 103–128, Apr. 2005.

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

[56] S. Grimme, "Semiempirical GGA-type density functional constructed with a long-range dispersion correction," *Journal of Computational Chemistry*, vol. 27, no. 15, pp. 1787–1799, 2006. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20495.

[57] 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," *Journal of Chemical Theory and Computation*, vol. 18, pp. 5410–5421, Sept. 2022. Publisher: American Chemical Society.

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

[59] C. Schran, K. Brezina, and O. Marsalek, "Committee neural network potentials control generalization errors and enable active learning," *The Journal of Chemical Physics*, vol. 153, p. 104105, Sept. 2020.

[60] K. Müller and L. D. Brown, "Location of saddle points and minimum energy paths by a constrained simplex optimization procedure," *Theoretica chimica acta*, vol. 53, pp. 75–93, Mar. 1979.

[61] A. France-Lanord, H. Vroylandt, M. Salanne, B. Rotenberg, A. M. Saitta, and F. Pietrucci, "Data-Driven Path Collective Variables," *Journal of Chemical Theory and Computation*, vol. 20, pp. 3069–3084, Apr. 2024. Publisher: American Chemical Society.

[62] Q. Li, B. Lin, and W. Ren, "Computing committor functions for the study of rare events using deep learning," *The Journal of Chemical Physics*, vol. 151, p. 054112, Aug. 2019.

[63] 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," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 15030–15035, Dec. 2015. Publisher: Proceedings of the National Academy of Sciences.

[64] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From A to B in free energy space," *The Journal of Chemical Physics*, vol. 126, p. 054103, Feb. 2007.

[65] R. Krishnamurthy, "Life's Biological Chemistry: A Destiny or Destination Starting from Prebiotic Chemistry?," *Chemistry – A European Journal*, vol. 24, no. 63, pp. 16708–16715, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/chem.201801847.

[66] J. Rogal, W. Lechner, J. Juraszek, B. Ensing, and P. G. Bolhuis, "The reweighted path ensemble," *The Journal of Chemical Physics*, vol. 133, p. 174109, Nov. 2010.

[67] B. Peters, "Reaction Coordinates and Mechanistic Hypothesis Tests," *Annual Review of Physical Chemistry*, vol. 67, no. 1, pp. 669–690, 2016. _eprint: https://doi.org/10.1146/annurev-physchem-040215-112215.

[68] A. Ma and A. R. Dinner, "Automatic Method for Identifying Reaction Coordinates in Complex Systems," *The Journal of Physical Chemistry B*, vol. 109, pp. 6769–6779, Apr. 2005. Publisher: American Chemical Society.

[69] A. P. de Alba Ortíz, J. Vreede, and B. Ensing, "The Adaptive Path Collective Variable: A Versatile Biasing Approach to Compute the Average Transition Path and Free Energy of Molecular Transitions," in *Biomolecular Simulations*, pp. 255–290, Springer, 2019.

[70] H. Jónsson, G. Mills, and K. W. Jacobsen, "Nudged elastic band method for finding minimum energy paths of transitions," in *Classical and Quantum Dynamics in Condensed Phase Simulations*, (LERICI, Villa Marigola), pp. 385–404, WORLD SCIENTIFIC, June 1998.

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

[72] A. Laio and F. L. Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Reports on Progress in Physics*, vol. 71, p. 126601, Nov. 2008.

[73] A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics," *WIREs Computational Molecular Science*, vol. 1, no. 5, pp. 826–843, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.31.

[74] C. Dellago and P. G. Bolhuis, "Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events," in *Advanced Computer Simulation Approaches for Soft Matter Sciences III* (C. Holm and K. Kremer, eds.), Advances in Polymer Science, pp. 167–233, Berlin, Heidelberg: Springer, 2009.

[75] P. G. Bolhuis and D. W. H. Swenson, "Transition Path Sampling as Markov Chain Monte Carlo of Trajectories: Recent Algorithms, Software, Applications, and Future Outlook," *Advanced Theory and Simulations*, vol. 4, no. 4, p. 2000237, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adts.202000237.

[76] B. Peters and B. L. Trout, "Obtaining reaction coordinates by likelihood maximization," *The Journal of Chemical Physics*, vol. 125, p. 054108, Aug. 2006.

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

[78] 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.

[79] 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.

[80] J. Kästner, "Umbrella sampling," *WIREs Computational Molecular Science*, vol. 1, no. 6, pp. 932–942, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.66.

[81] 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.

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

[83] 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.

[84] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992. Publisher: Wiley Online Library.

[85] J. Perrin, "Mouvement brownien et réalité moléculaire," *Annales de chimie et de physique*, 1909.

[86] A. Einstein, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," *Annalen der Physik*, vol. 322, no. 8, pp. 549–560, 1905. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19053220806.

[87] L. Boltzmann, *Theoretical Physics and Philosophical Problems*. Dordrecht: Springer Netherlands, 1974.

[88] D. S. Lemons and A. Gythiel, "Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) **146** , 530–533 (1908)]," *American Journal of Physics*, vol. 65, pp. 1079–1081, Nov. 1997.

[89] D. D. Girardier, H. Vroylandt, S. Bonella, and F. Pietrucci, "Inferring free-energy barriers and kinetic rates from molecular dynamics via underdamped Langevin models," *The Journal of Chemical Physics*, vol. 159, p. 164111, Oct. 2023.

[90] A. Nitzan, *Chemical dynamics in condensed phases: relaxation, transfer and reactions in condensed molecular systems*. Oxford graduate texts, Oxford ; New York: Oxford University Press, 2006. OCLC: ocm62118341.

[91] H. Vroylandt, "On the derivation of the generalized Langevin equation and the fluctuation-dissipation theorem," *Europhysics Letters*, vol. 140, p. 62003, Dec. 2022. Publisher: EDP Sciences, IOP Publishing and Società Italiana di Fisica.

[92] H. B. Callen and T. A. Welton, "Irreversibility and Generalized Noise," *Physical Review*, vol. 83, pp. 34–40, July 1951. Publisher: American Physical Society.

[93] H. A. Kramers, "Brownian motion in a field of force and the diffusion model of chemical reactions," *Physica*, vol. 7, pp. 284–304, Apr. 1940.

[94] M. A. Islam, "Einstein–Smoluchowski Diffusion Equation: A Discussion," *Physica Scripta*, vol. 70, p. 120, Jan. 2004.

[95] S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy," *Reviews of Modern Physics*, vol. 15, pp. 1–89, Jan. 1943.

[96] H. Vroylandt and P. Monmarché, "Position-dependent memory kernel in generalized Langevin equations: Theory and numerical estimation," *The Journal of Chemical Physics*, vol. 156, p. 244105, June 2022.

[97] M. Ceriotti, G. Bussi, and M. Parrinello, "Colored-Noise Thermostats à la Carte," *Journal of Chemical Theory and Computation*, vol. 6, pp. 1170–1180, Apr. 2010. Publisher: American Chemical Society.

[98] H. Vroylandt, L. Goudenège, P. Monmarché, F. Pietrucci, and B. Rotenberg, "Likelihood-based non-Markovian models from molecular dynamics," *Proceedings of the National Academy of Sciences*, vol. 119, p. e2117586119, Mar. 2022. arXiv:2110.04246 [cond-mat, physics:physics].

[99] K. Palacio-Rodriguez, H. Vroylandt, L. S. Stelzl, F. Pietrucci, G. Hummer, and P. Cossio, "Transition Rates and Efficiency of Collective Variables from Time-Dependent Biased Simulations," *The Journal of Physical Chemistry Letters*, vol. 13, pp. 7490–7496, Aug. 2022. Publisher: American Chemical Society.

[100] L. Mouaffac, K. Palacio-Rodriguez, and F. Pietrucci, "Optimal Reaction Coordinates and Kinetic Rates from the Projected Dynamics of Transition Paths," *Journal of Chemical Theory and Computation*, vol. 19, pp. 5701–5711, Sept. 2023. Publisher: American Chemical Society.

[101] S. Ghysbrecht, L. Donati, and B. G. Keller, "Accuracy of reaction coordinate based rate theories for modelling chemical reactions: insights from the thermal isomerization in retinal," Dec. 2023. arXiv:2312.12948 [physics].

[102] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, "Mori–Zwanzig formalism as a practical computational tool," *Faraday Discussions*, vol. 144, no. 0, pp. 301–322, 2010. Publisher: Royal Society of Chemistry.

[103] K. J. Laidler, "A glossary of terms used in chemical kinetics, including reaction dynamics (IUPAC Recommendations 1996)," *Pure and Applied Chemistry*, vol. 68, pp. 149–192, Jan. 1996. Publisher: De Gruyter.

[104] C. H. BENNETT, "Molecular Dynamics and Transition State Theory: The Simulation of Infrequent Events," in *Algorithms for Chemical Computations*, vol. 46 of *ACS Symposium Series*, pp. 63–97, AMERICAN CHEMICAL SOCIETY, June 1977. Section: 4.

[105] D. Chandler, "Statistical mechanics of isomerization dynamics in liquids and the transition state approximation," *The Journal of Chemical Physics*, vol. 68, pp. 2959–2970, Aug. 2008.

[106] K. Palacio-Rodriguez and F. Pietrucci, "Free Energy Landscapes, Diffusion Coefficients, and Kinetic Rates from Transition Paths," *Journal of Chemical Theory and Computation*, vol. 18, pp. 4639–4648, Aug. 2022. Publisher: American Chemical Society.

[107] J. O. Daldrop, B. G. Kowalik, and R. R. Netz, "External Potential Modifies Friction of Molecular Solutes in Water," *Physical Review X*, vol. 7, p. 041065, Dec. 2017.

Glossary

AIMD *Ab Initio* Molecular Dynamics. 25, 34, 35, 36, 37, 46, 47, 54, 56, 58, 64, 66, 92, 151, 154

AS Aimless Shooting. 52, 54

BOMD Born-Oppenheimer Molecular Dynamics. 26

CA Committor Analysis. 47, 48, 49, 52

CV Collective Variable. 24, 36, 39, 40, 41, 44, 47, 49, 50, 51, 52, 54, 56, 58

DFT Density Functional Theory. 29, 31, 33, 34, 35, 37, 43

GGA Generalized Gradient Approximation. 32

GLE Generalized Langevin Equation. 62, 153, 154

GPW Gaussian Plane Wave. 34

LDA Local Density Approximation. 31, 32

MFEP Minimum Free Energy Path. 40, 52, 54, 55, 68, 154

MFPT Mean First Passage Time. 64, 65

MLIP Machine Learning Inter-atomic Potential. 33, 35, 36, 37, 39

MLMD Machine Learned Molecular Dynamics. 35, 36, 37, 92, 154

NEB Nudged Elastic Band. 48

NHC Nose-Hoover Chain. 23

NNP Neural Network Potential. 36, 37, 92

OP Order Parameter. 40, 46, 47, 54, 65

PBC Periodic Boundary Conditions. 20, 34

PBE Perdew-Burke-Ernzerhof. 31, 32, 33

PCV Path Collective Variable. 44, 45, 46, 48, 50, 55, 58, 91, 155

PW Plane Wave. 34, 35

RC Reaction Coordinate. 40, 47, 49, 56, 58, 64, 91

SCF Self-Consistent Field. 27, 29, 31

SFT Shooting From the Top. 54, 55

SLLS Standard Log-Likelihood Score. 91

TPS Transition Path Sampling. 52, 54, 66

TS Transition State. 47, 48, 52, 54

ULE Underdamped Langevin Equation. 155

US Umbrella Sampling. 56, 57, 58, 66, 92

VDW van der Waals. 35

WHAM Weighted Histogram Analysis Method. 56, 58

Appendices

Appendix A

Dissemination of research results and teaching activities

A.1 Publications

A.1.1 Published papers

- 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 2022, 126, 47, 8887–8900

A.1.2 Accepted papers

- L. Huet, T. Devergne, F. Pietrucci and A. M. Saitta "A new route to the prebiotic synthesis of glycine via quantum-based machine learning calculations" *Accepted with minor revisions to J. Phys. Chem. L.*

- L. Huet, T. Magrino, A. M. Saitta and F. Pietrucci, "Correction to 'Step by Step Strecker Amino Acid Synthesis from ab Initio Prebiotic Chemistry'" *Accepted to J. Phys. Chem. L.*

- 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", *Accepted to Phys. Rev. Let.*

A.1.3 Papers in preparation

- L. Huet, H. Voylandt, R. Vuilleumier, A. M. Saitta and F. Pietrucci "Insight on chemical reaction dynamics and reaction coordinates from non-Markovian models", *In preparation*

A.2 Participation to conferences

A.2.1 Contributed talk

- Stochastic models form enhanced sampling in Chemistry, *Probabilistic Sampling For Physics*, 04/09/2023-22/09/2023, Orsay, France
- Agnostic exploration of a new pathway for prebiotic glycine synthesis by ab initio and machine learning molecular dynamics, *AbSciCon*, 05/05/2024-10/05/2024, Providence, Rhode Island, USA
- Elucidating Novel Pathways for Prebiotic Glycine Synthesis: Merging Enhanced Sampling and Neural Network Atomic Potentials, *ThéMoSiA - RCTF2024*, 24/06/2024-28/06/2024, Rouen, France
- Agnostic exploration of a new pathway for prebiotic glycine synthesis by ab initio and machine learning molecular dynamic *Journées "Théorie, Modélisation et Simulation*, 17/10/2023-18/10/2023, Strasbourg, France

A.2.2 Posters

- Prebiotic chemistry pathway discovered by a fully agnostic ab initio molecular dynamics method, *Exobiologie Jeunes Chercheurs*, 17/10/2022-18/10/2022, Paris, France
- A prebiotic chemistry pathway discovered by a fully agnostic ab-initio molecular dynamics method, *BEACON*, 04/05/2023-14/05/2023, La Palma Island, Canary Islands, Spain
- Ab initio and machine learned molecular dynamic for prebiotic chemistry thermodynamics and kinetics, *Journées plénières 2024 du GDR IAMAT*, 02/07/2024-05/07/2024, Toulouse, France

A.3 Teaching activities

- Projet informatique, M1 level at Chimie ParisTech (September to December 2023) (60h TP)

Appendix B

Supporting information of "Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry"

Supporting Information: Critical Assessment of Data-Driven *versus* Heuristic Reaction Coordinates in Solution Chemistry

Théo Magrino,[†,‡] Léon Huet,[†,‡] A. Marco Saitta,[†] and Fabio Pietrucci[*,†]

†*Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS UMR 7590, Paris 75005 FRANCE*

‡*These two authors contributed equally*

E-mail: fabio.pietrucci@sorbonne-universite.fr

Modified shooting-from-the-top algorithm

The algorithm employed to perform shooting from the top was very close to the one originally proposed in ref.,[1] however, to generate a maximum number of trajectories for the definition of path CVs we set the acceptance probability to 1. The following analysis shows that this modification does not lead, in our specific system, to a sizable difference in relevant physical properties with respect to the original algorithm.

Following ref.,[1] we denote X the last accepted pathway and X' the new proposed pathway. Probability of pathways X and X' are $p[X]$ and $p[X']$ and they contain respectively n and n' points in the CV range used for shooting from the top (*i.e.* $s_2 \in [1.35, 1.65]$ for $SFT_1.35 - 1.65$).

Denoting $p_{gen}[X \to X']$ and $p_{acc}[X \to X']$ the probabilities of generating and accepting pathway X' from X, the following condition must be satisfied to obtain unbiased sampling:

$$\frac{p[X]}{p[X']} = \frac{p_{gen}[X \to X'] \times p_{acc}[X \to X']}{p_{gen}[X' \to X] \times p_{acc}[X' \to X]} \tag{1}$$

(see expression (3) in ref.[1]). Equations (6), (7) and (8) from ref.[1] show that:

$$\frac{p_{acc}[X \to X']}{p_{acc}[X' \to X]} = \frac{n}{n'} \tag{2}$$

However, in the present work, we used a simpler approach, corresponding to biased acceptances $p^B_{acc} = 1$ independent from the path length in the the CV range. The resulting biased probabilities $p^B[X]$ and $p^B[X']$ therefore obey a similar equation as equation (1) with acceptance ratio equal to 1:

$$\frac{p^B[X]}{p^B[X']} = \frac{p_{gen}[X \to X']}{p_{gen}[X' \to X]} \times 1 \tag{3}$$

Combining (4), (1) and (2) :

$$\frac{p[X]/n}{p[X']/n'} = \frac{p^B[X]}{p^B[X']} \tag{4}$$

we therefore disfavor pathways spending longer time in the CV shooting range.

We performed a comparison of the original and modified path probabilities $p[X]$ and $p^B[X]$ by comparing the latter with the former, obtained via reweighting $p^B[X] \to p^B[X] \cdot n$ (followed by normalization), that increases the relative weight of "long" reactive paths.

The average length of sampled paths is 62 fs (standard deviation 27 fs), while after reweighting the average length is 74 fs (standard deviation 32 fs), thus the two distributions have a sizable overlap. We also checked the effect of the modified algorithm by performeing the same PCA analysis of CVs directly on the sampled data and after reweighting. Results are presented in fig. S1, showing the lack of sizable differences.

These results could be anticipated, since the n/n' acceptance ratio shows non-trivial behavior mostly for diffusive barriers that can feature a wide range of crossing times, as in protein conformational changes. In the case of chemical reactions, narrow barriers and lower-friction dynamics is expected to lead to narrower crossing-times distribution.

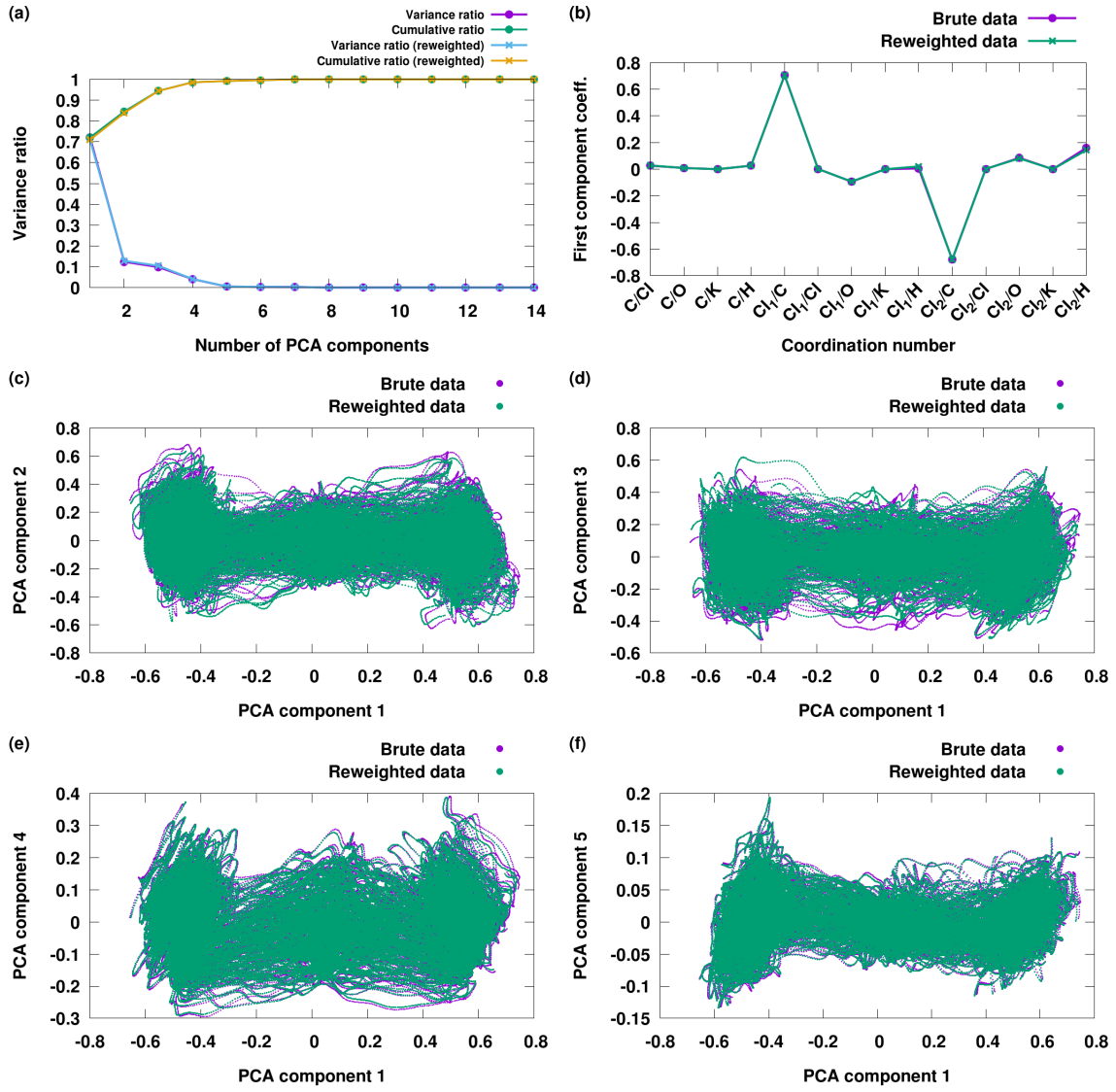


Figure S1: Evaluation of the effect of reweighting shooting-from-the-top path probabilities on PCA results. **(a)** Variance ratio for PCA components and cumulative variance ratio. **(b)** Coefficients in terms of the data (coordination numbers) of the first PCA component ($\sim$ 72% of the total variance). **(c)**, **(d)**, **(e)** and **(f)**: representations of data in terms of PCA component 1 and, respectively, PCA components 2, 3, 4, 5 (explaining 99% of the variance).

PCA and clustering of the transition state ensemble

We analyzed the ensemble of accepted shooting points from $SFT_{1.35-1.65}$ data set A in the space of coordination number CVs.

PCA results are presented in figure S2, after the projecting the space of coordination numbers on the four principal components.

Cluster analysis in the TS ensemble was performed to check whether the latter contains a significant structural diversity, with possible multiple mechanistic channels. To this aim, we employed the k-means and Gaussian mixture model algorithms as implemented in scikit-learn.[2] All parameters were set to default except the number of initializations n_init, that we increased to 50 (default value 1). We considered a number of clusters (respectively, Gaussian components) varying from 2 to 30. We estimated the statistical dispersion of results from 30 repetitions of the clustering. We then evaluated results with:

(i) Cluster inertia for k-means algorithm (fig. S3 (a)/(d)).

(ii) Silhouette scores for both k-means and Gaussian mixture algorithms (fig. S3 (b)/(e)).

(iii) For Gaussian mixture algorithm, we computed the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) (fig. S3 (c)/(f)).

To evaluate the optimal number of clusters, one may seek (i) an elbow in inertia plots (a) and (d), (ii) a maximal value of silhouette score or (iii) a minimal value for AIC and BIC. Such features cannot be identified before PCA (panels (a), (b) and (c)). After PCA the same conclusion holds for k-means clustering (panel (d)). Assuming a maximal value for Gaussian mixture after PCA silhouette score for 8 Gaussian components (panel (e)), this would not be consistent with the small AIC minimum for 4 Gaussian components (panel (f)). We therefore conclude that no significant cluster structure could be deduced for the $SFT_{1.35-1.65}$ TS ensemble, pointing to a structurally homogeneous ensemble and to a single reaction mechanism.

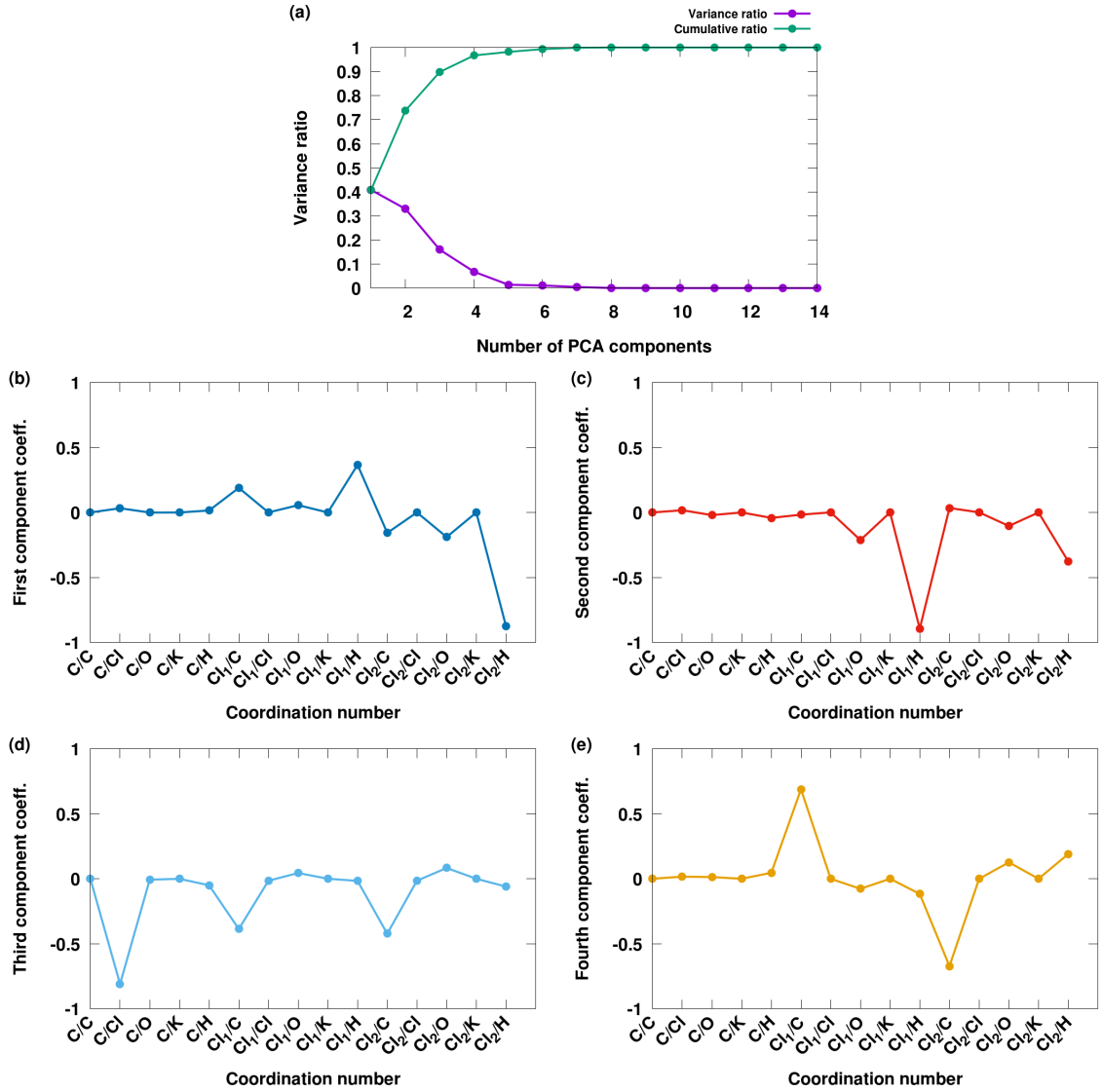


Figure S2: PCA of accepted shooting points from $SFT_{1.35-1.65}$ data set A. **(a)**: fraction of variance explained and cumulative variance for successive components. **(b,c,d,e)**: coefficient of the first to fourth component as a function of coordination number CVs.

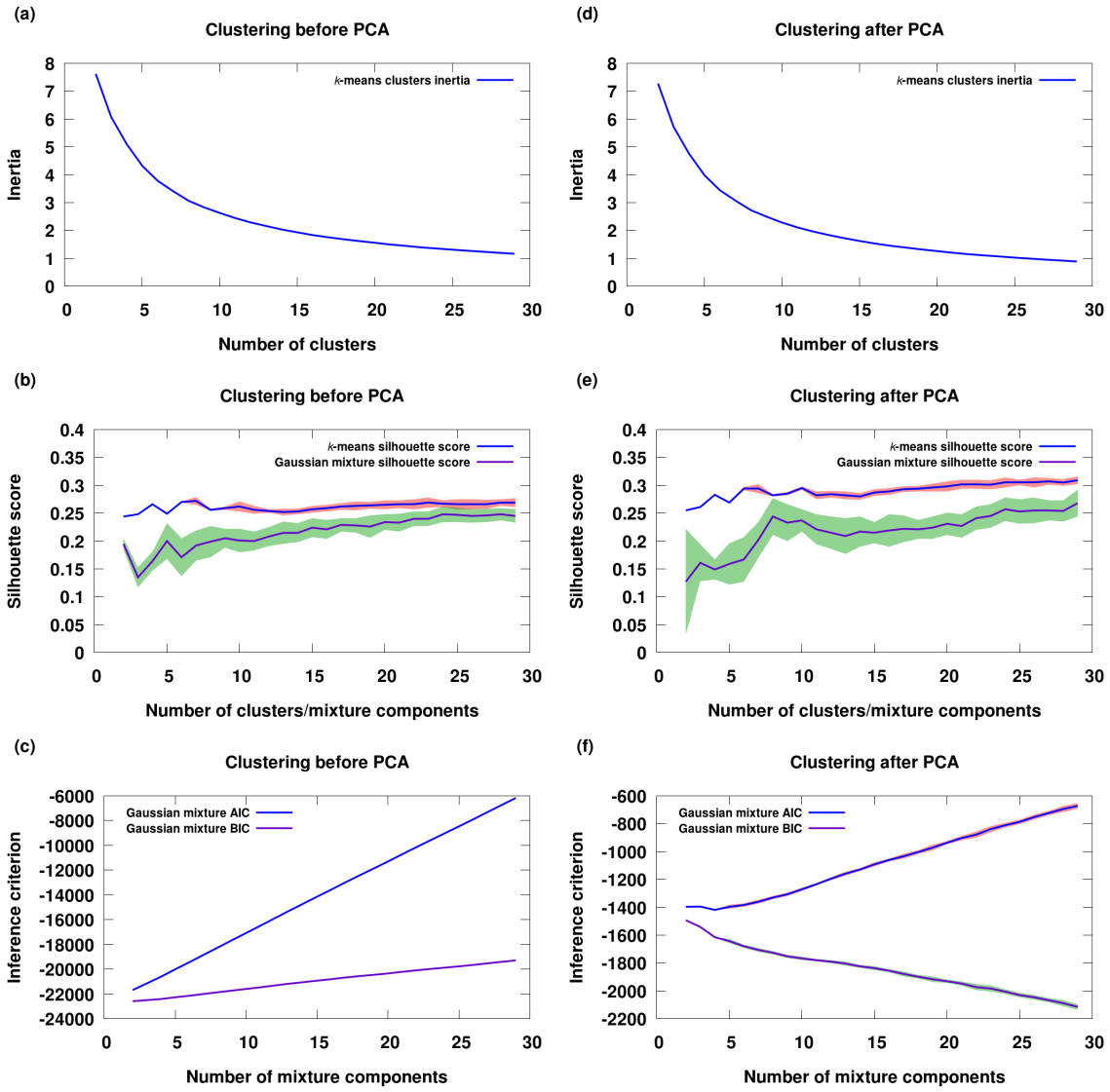


Figure S3: $SFT_{1.35-1.65}$ clustering information. **(a)**: Clusters inertia as a function of the number of k-means cluster before PCA. **(b)**: Silhouette scores for k-means clusters and Gaussian mixture components before PCA. **(c)** Akaike and Bayesian information criteria before PCA. **(d)**: Clusters inertia as a function of the number of k-means cluster after PCA. **(e)**: Silhouette scores for k-means clusters and Gaussian mixture components after PCA. **(f)** Akaike and Bayesian information criteria after PCA. Background colors represent maximal and minimal value for each of the 30 replicas done for each algorithm version. They are clearly visible only on panels (b) and (e).

Optimized coordinates based on SLLS

Table S1: Intervals of sampled values for the CVs used for PCA and SLLS.

CV	min	max	Δ
c_1	0.0975	0.871	0.774
c_2	0.117	0.902	0.785
c(Cl_1,H)	0.928	1.66	0.727
c(Cl_2,H)	0.928	1.77	0.84
c(C,O)	0.136	0.185	0.0488
c(C,K)	$3.83 \cdot 10^{-5}$	$7.68 \cdot 10^{-5}$	$3.85 \cdot 10^{-5}$
c(C,H)	2.89	3.04	0.153
c(Cl_1,Cl_2)	0.00578	0.0142	0.00841
c(Cl_1,O)	0.544	0.716	0.172
c(Cl_1,K)	$4.19 \cdot 10^{-5}$	$6.59 \cdot 10^{-5}$	$2.41 \cdot 10^{-5}$
c(Cl_2,O)	0.538	0.812	0.273
c(Cl_2,K)	$1.41 \cdot 10^{-4}$	$2.08 \cdot 10^{-4}$	$6.76 \cdot 10^{-5}$
rand	$1.10 \cdot 10^{-4}$	0.999	0.999

Table S2: Complete table of the optimized coordinates. We stop at oc_7: the presence of the random CV in the linear combination suggests that no more information is needed as *rand* is completely decorrelated form the reaction by definition. All the coefficients are from the first set of data.

oc_x	Colvars	Formula
oc_1	c_1	$4.9 - 9.6c_1$
oc_2	c_1 ; c(Cl_1,Cl_2)	$2.2 - 1.0 \cdot 10c_1 + 2.0 \cdot 10^2$c($Cl_1$,$Cl_2$)
oc_{c_1,c_2}	c_1 ; c_2	$0.6 - 5.3c_1 + 4.7c_2$
oc_3	c_1 ; c_2 ; c(Cl_1,O)	$-9.4 - 3.3c_1 + 1.0 \cdot 10c_2 + 1.1 \cdot 10$c($Cl_1$,O)
oc_4	c_2 ; c(Cl_2,H) ; c(Cl,Cl) ; c(Cl_1,O)	$-9.7 + 1.4 \cdot 10c_2 - 1.4$c($Cl_2$,H)$-1.2 \cdot 10^2$c($Cl_1$,$Cl_2$) $+1.1 \cdot 10$c(Cl_1,O)
oc_5	c_2; c(C,H); c(Cl_1,Cl_2); c(Cl_1,O); c(Cl_2,K)	$2.2 \cdot 10 + 1.0 \cdot 10c_2 - 7.1$c(C,H)$-2.2 \cdot 10^2$c($Cl_1$,$Cl_2$) + 8.6c(Cl_1,O)$-4.5 \cdot 10^4$c(Cl_2,K)
oc_6	c_2; c(Cl_2H); c(C,O); c(C,H); c(Cl_1,Cl_2); c(Cl_1,O)	$1.4 \cdot 10 + 1.3 \cdot 10c_2 - 3.0$c($Cl_2$,H)$+2.2 \cdot$ 10c(C,O)-7.2c(C,H)$-2.5 \cdot 10^2$c(Cl_1,Cl_2) + 7.5c(Cl_1,O)
oc_7	c_2; c(Cl_2,H); c(C,H); c(Cl_1,Cl_2); c(Cl_1,O); c(Cl_2,K); rand	$-6.8 + 1.4 \cdot 10c_2 - 1.4$c($Cl_2$,H)$-9.8 \cdot 10^{-2}$c(C,H)$-1.8 \cdot$ 10^2c(Cl_1,Cl_2) + $1.1 \cdot 10$c(Cl_1,O)$-1.1 \cdot 10^4$c(Cl_2,K) $-3.5 \cdot 10^{-1}$rand

Umbrella sampling configuration space projected on non-biased CVs

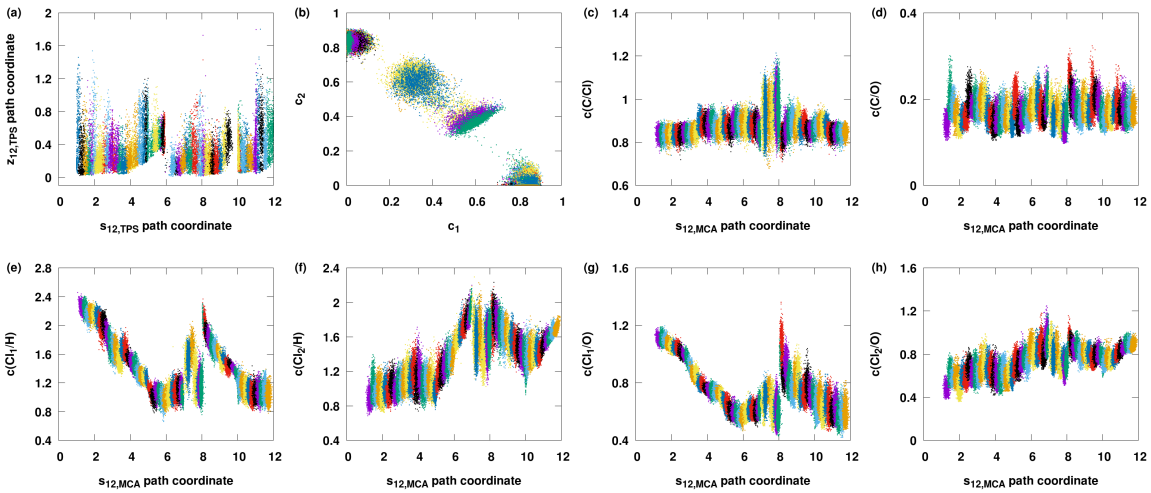


Figure S4: US trajectories obtained biasing $s_{12,MCA}$ (no additional restraints) projected along different CVs. Only the second half of simulations is represented (from 9 to 18 ps for each US window).

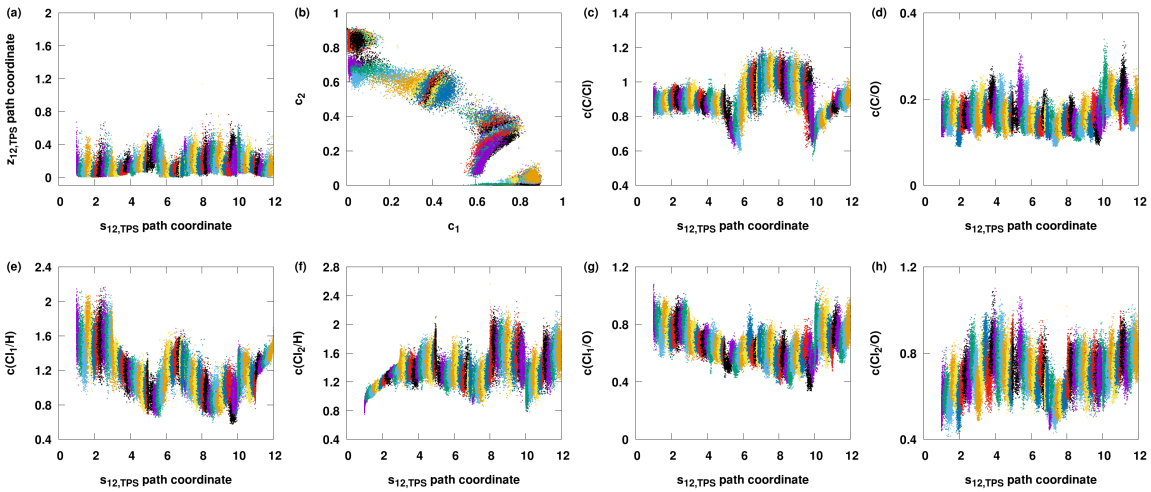


Figure S5: US trajectories obtained biasing $s_{12,TPS}$ (no additional restraints) projected along different CVs. Only the second half of simulations is represented (from 9 to 19 ps for each US window).

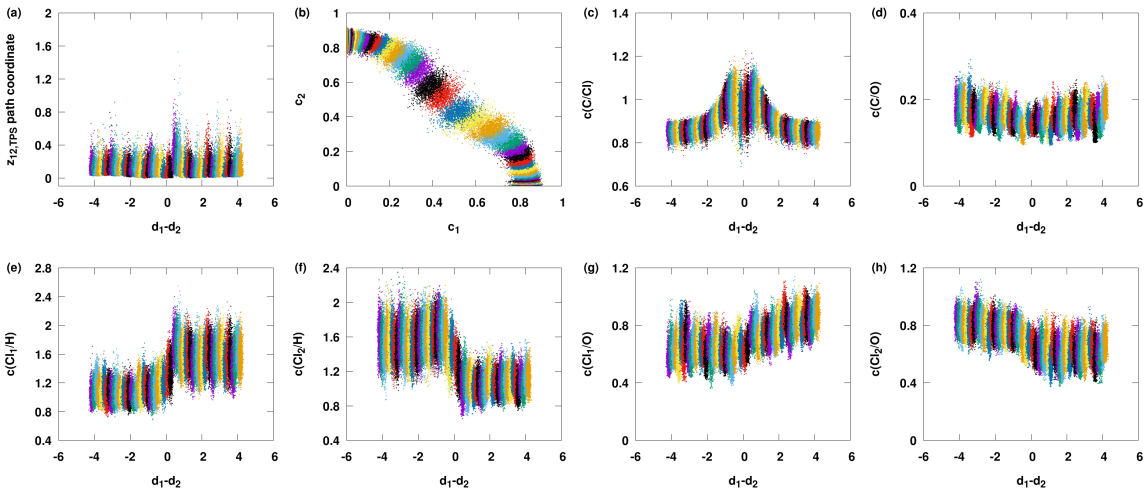


Figure S6: US trajectories obtained biasing $d_1 - d_2$ (no additional restraints) projected along different CVs. Only the second half of simulations is represented (from 10 to 21 ps for each US window).

Complementary information on US

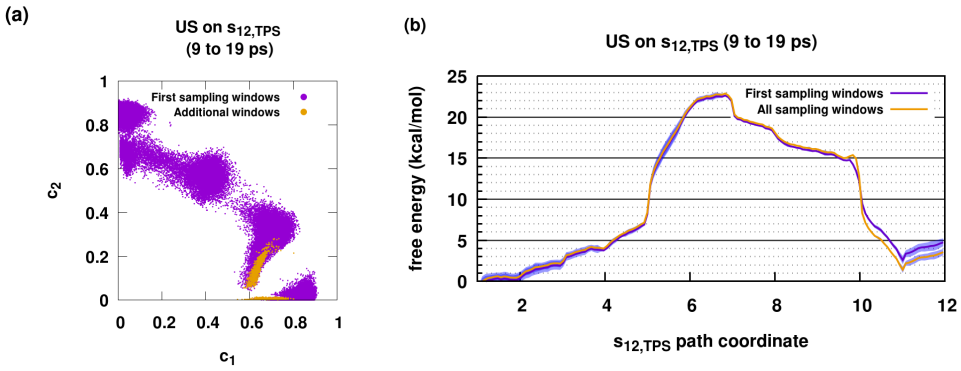


Figure S7: Effect of a sampling gap in the (c_1,c_2) plane (panel **(a)**), absent in the ($s_{12,\mathrm{TPS}}$,$z_{12,\mathrm{TPS}}$) plane, on the US free-energy landscape reconstructed with WHAM ((panel **(b)**). The orange landscape includes additional configurations sampled close to the gap (orange points in **(a)**). Statistical uncertainties are indicated in pale blue.

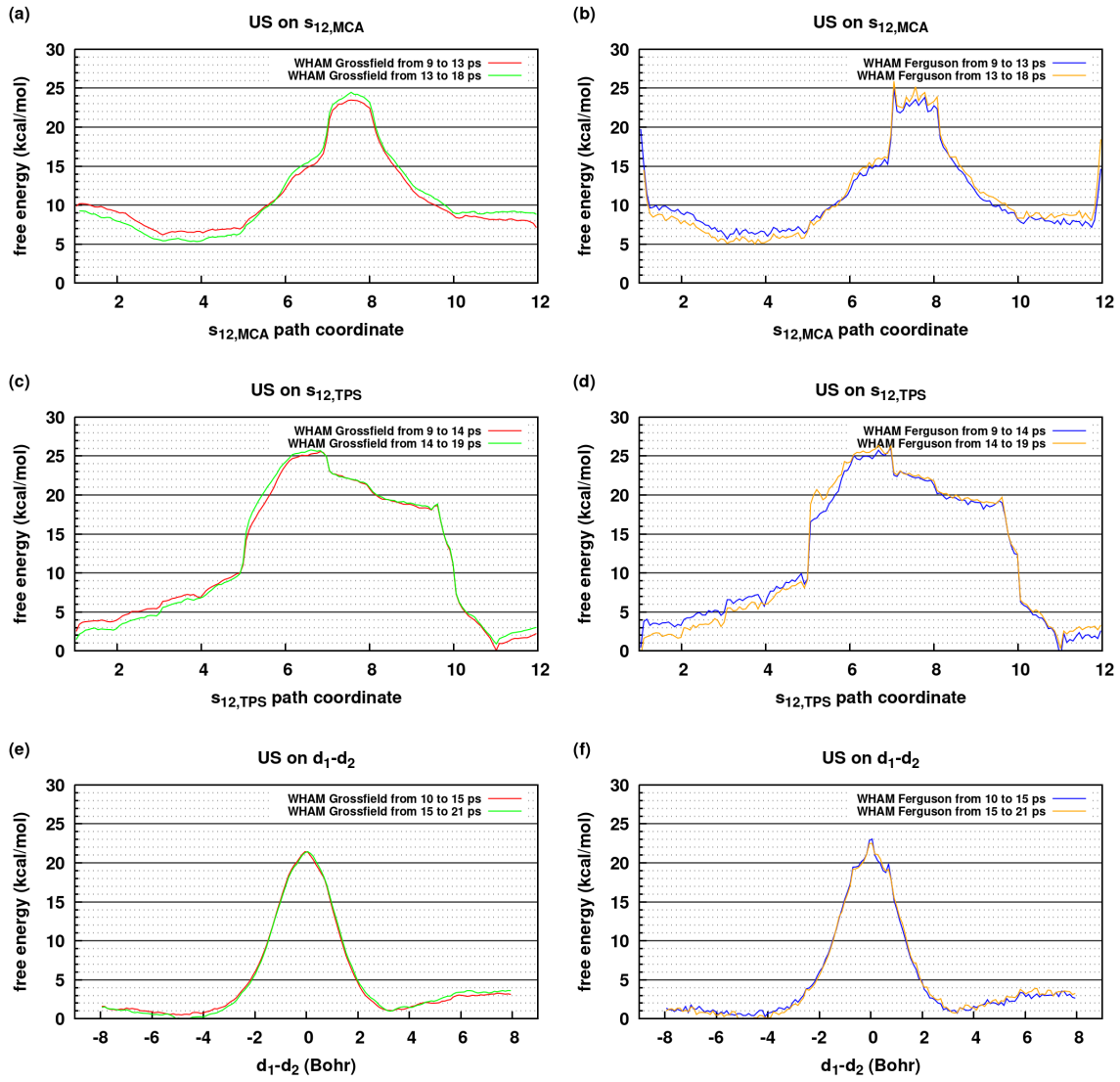


Figure S8: Comparison of free-energy landscapes reconstructed from US simulations using two different WHAM codes, from refs.[3] and,[4] respectively.

References

(1) Jung, H.; Okazaki, K.-i.; Hummer, G. Transition path sampling of rare events by shooting from the top. *The Journal of chemical physics* **2017**, *147*, 152716.

(2) Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(3) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.11. `http://membrane.urmc.rochester.edu/?page_id=126`, Last accessed 28 September 2022.

(4) Ferguson, A. L. BayesWHAM: A Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. *Journal of Computational Chemistry* **2017**, *38*, 1583–1605, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.24800.

Appendix C

Supporting information of "A new route to the prebiotic synthesis of glycine via quantum-based machine learning calculations"

Supporting information for "A new route to the prebiotic synthesis of glycine via quantum-based machine learning calculations"

Léon Huet,[†] Timothée Devergne,[†,‡,¶] Théo Magrino,[†] and A. Marco SAITTA[*,†]

†*Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS UMR 7590, Paris 75005, FRANCE*

‡*Atomistic Simulations, Italian Institute of Technology, 16142 Genoa, Italy*

¶*Computational Statistics and Machine Learning, Italian Institute of Technology, 16142 Genoa, Italy*

E-mail: marco.saitta@sorbonne-universite.fr

Phone: +33 1 44 27 22 44. Fax: +33 1 44 27 51 52

Computational protocol details

In this section we explore further details about the protocol that are not mentioned in the paper.

Input template

Here is our CPMD input template:

```
&CPMD
  MOLECULAR DYNAMICS BO
```

```
CONVERGENCE ORBITALS
 1.e-5
reSTART COORDINATES WAVEFUNCTION LATEST VELOCITIES
VDW CORRECTION
STORE WAVEFUNCTIONS
 100
EXTRAPOLATE WFN
 4
TIMESTEP
 20.
TEMPERATURE
 300.
NOSE IONS MASSIVE
 300. 3000.
TRAJECTORY XYZ SAMPLE FORCES
 1
MAXSTEP
 100000
MAXRUNTIME
 3500
COMPRESS WRITE32
MIRROR
RESTFILE
 2
MEMORY BIG
REAL SPACE WFN KEEP
INITIALIZE WAVEFUNCTION RANDOM
ALLTOALL SINGLE
CP_GROUPS
```

```
   1
&END

&DFT
  FUNCTIONAL PBE
  GC-CUTOFF
   1.E-05
&END

&VDW
 EMPIRICAL CORRECTION
 VDW PARAMETERS
 ALL DFT-D2
 S6GRIM
 PBE
 END EMPIRICAL CORRECTION
&END

&SYSTEM
  ANGSTROM
  SYMMETRY
   1
  CELL
   13.36074 1 1 0  0  0
  CUTOFF
   80
&END

&ATOMS
```

```
*C_MT_PBE  KLEINMAN-BYLANDER
 LMAX=P
CCC
*O_MT_PBE  KLEINMAN-BYLANDER
 LMAX=P
OOO
*N_MT_PBE  KLEINMAN-BYLANDER
 LMAX=P
NNN
*H_MT_PBE  KLEINMAN-BYLANDER
 LMAX=S
HHH
 ISOTOPES
  12.0107
  15.9994
  14.0067
   2.0
&END
```

To increase the stability of the model the mass of H elements has been set to 2 a.u., modeling heavy water. This diminish the O-H covalent bond oscillation frequency,[1,2] and increase also the reliability of the Born Oppenheimer approximation,[2] ensuring numerical stability.

"CCC", "OOO", "NNN" and "HHH" denote places where the number of each element in the box are inserted followed by the xyz coordinates of the initial geometry.

Path Collective Variables: A more Mathematic point of view

PCVs have been established in 2007 by Branduardi *et al.*[3] The idea is to start from a set of reference vectors composed of collective variables: $\{R^i\}$, sorted to pin a path of configurations between reactants and products. The collective variables inside each R^i are selected to grasp

the essential degrees of freedom of the reaction. Then the two PCVs, s and z, can be defined as in the equation 1, where s represents the progress along the pined reference path and z the general distance with all the references:

$$\begin{cases} s(\mathrm{t}) & = \dfrac{\sum_{\alpha=1}^{N} \alpha \times \exp(-\lambda \mathrm{D}[\mathrm{R}_{x(t)}, \mathrm{R}^{\alpha}])}{\sum_{\alpha=1}^{N} \exp(-\lambda \mathrm{D}[\mathrm{R}_{x(t)}, \mathrm{R}^{\alpha}])} \\ \\ z(\mathrm{t}) & = \dfrac{-1}{\lambda} \ \ln\left(\sum_{\alpha=1}^{N} \exp(-\lambda \mathrm{D}[\mathrm{R}_{x(t)}, \mathrm{R}^{\alpha}])\right) \end{cases} \tag{1}$$

In equation 1, D represented a distance metric in the vector space of our chosen set of collective variables, making of it a metric space. R^{α} is the αth reference vector and $\mathrm{R}_{x(t)}$ is the collective variable vector of the geometry $x(t)$. λ is an external parameter that we define using the equation 2, in order to guarantee a smooth behaviour of s and z, as in our references.[4–6] Too large values of λ can lead a discontinuous behaviour of s and z whereas too small values blur the separation between references making intermediate geometries undistinguished.

$$\lambda \times mean(\mathrm{D}[\mathrm{R}^{\alpha}, \mathrm{R}^{\alpha+1}]) = -ln(0.1) \approx 2.30 \tag{2}$$

The distance metric used in this work is taken from the same previous works:.[4–6] It is defined as the square of the euclidean distance:

$$\mathrm{D}[\mathrm{R}_{x(t)}, \mathrm{R}^{\alpha}] = \sum_{i \in \mathrm{A}} \sum_{\sigma \in \mathrm{E}} [c_{i\sigma}(x(t)) - c_{i\sigma}^{\alpha}]^2 \tag{3}$$

Where:

- A is a set of atom, that we chose to follow
- E is the set of different periodic elements in the system
- $c_{i\sigma}(x(t))$ is the coordinance, in element σ, of the atom i, for the geometry x(t)

- $c_{i\sigma}^{\alpha}$ the $i\sigma$ component of reference vector R^{α} corresponding to the targeted coordinance.

The s(t) variable of the PCVs behaves in a very intuitive way as a reaction coordinate, with a two times in a raw reduction of the dimensionality of the problem. The first one consists of passing from the 3N phase-space geometry to a reference vector of roughly 10 to 20 dimensions, which should condense the definition of the chemical states associated with each geometry. Our choice to use the coordinance of a set of atoms to reduce the dimensionality of the problem proved its reliability in our previous works, specifically in the case of covalent cleaving and bonding in solution.[4,5,7–10] The second reduction of dimension is the calculation of the one-dimensional s and z variables.

To permit comparison between the different steps of the mechanism and the different steps of the protocol, each s variable has been normalised to evolve in the $[0, 1]$ interval (0 for reactants, 1 for products), instead of the $[1, N]$ interval of its definition, by a linear rescaling.

Atomic permutations in PCVs

The reference Matrices that are used to define PCVs are not specific to a chemical state. Among these differences a chemical state is the same if there are permutations between atomic positions of the same element. For example, if the two carbons of the glycine are inverted, then the molecule is still the same from a chemical and physical point of view. That is not a dilemma if we know the mechanism that is going to occur in the simulation box, but in the case of an explorative metadynamics we do not know which carbon of the reactants will correspond to which one in the products. We have to make a choice. As with any human choice in our protocol, we want to point this out as it is a possible source of errors. In the case of glycine synthesis, as there is already one well known synthesis pathway, the Strecker one, we decided to sort the products lines to match with it. In general case, we propose to choose among all the possible permutations of the products' reference matrix the one that is the closest to reactants in matrix distance, (i-e the one that should imply

fewer modifications). This reference could be named the Least Variation Permuted Reference (LVPR). If chemical arguments are used against the LVPR, like in our case with the streaker mechanism, then man can use another permutation.

	Elements σ			
	C	O	N	H
C^1	0.0	1.2	0.0	2.0
C^2	0.0	0.2	1.0	1.1
N^1	0.0	0.3	0.0	3.4
N^2	1.0	0.2	0.0	0.5

(a) Reactants reference table for every mechanism

	Elements σ			
	C	O	N	H
C^1	0.8	0.5	0.9	2.3
C^2	0.8	2.0	0.1	0.6
N^1	1.0	0.3	0.0	2.6
N^2	0.0	0.3	0.0	3.4

(b) Products reference table from Strecker mechanism

	Elements σ			
	C	O	N	H
C^2	0.8	2.0	0.1	0.6
C^1	0.8	0.5	0.9	2.3
N^2	0.0	0.3	0.0	3.4
N^1	1.0	0.3	0.0	2.6

(c) Products LVM

Figure S-1: Least variation permuted reference, compared with Strecker reference for products.

For the oxyglycolate pass toward glycine, the final geometry we obtain is glycine, but is not corresponding exactly to the reference matrix R^2 we used. At the final state position of N^1 and N^2 are permuted, leading to the same chemical system but to another coordination matrix. It is observable in the metadynamic bias footprint figure in the article where the position of the reactants and products wells are not symmetric.

Bonding distance

$$c_i(\sigma) = \sum_{j \in \sigma} \frac{1 - \left(\frac{d_{ij}}{d_0}\right)^8}{1 - \left(\frac{d_{ij}}{d_0}\right)^{14}} \tag{4}$$

The smooth switching function, see figure S-2, is used to calculate the coordination of an atom for one atomic element using the equation 4 The d_0 in the equation is a cutoff distance of bonding specific to the studied couple. In our calculation we used the value stated in table S-1.

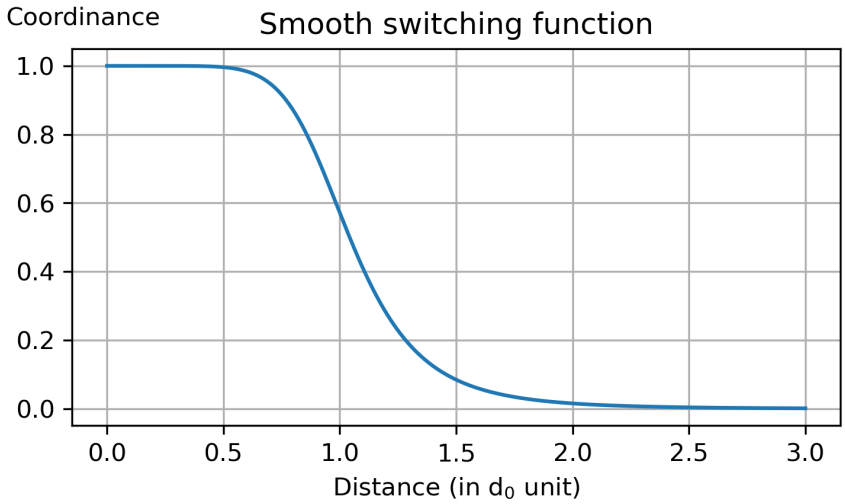


Figure S-2: Graph of the smooth switching function in units of d_0. Every atom at d_0 distance will be counted with a coordinate of +0.6 for the concerned element. The total coordination of an atom for an element is the sum of all the smooth switching functions for all the species of this element in the simulation simulation box.

Table S-1: Switching function separation distance

Atom	Element	d_0	comments
X	W	1.80Å	X≠H, W≠H
C	H	1.20Å	
N	H	1.10Å	
O	H	1.10Å	

As switching functions are not directly used to make physical measurements out of our work, the value of d_0 has not to be precisely equal to the covalent bonding distance but has to be set to separate correctly bonded interactions (near 1) to non-bonded interaction (less than 0.5). This is done by mapping the histogram of distance of atom with an element in the bulk. The value of 1.1Åfor O-H and N-H is set to correctly discriminate between covalent bonds and H-bonds

Umbrella sampling: standard errors and check protocol

Umbrella Sampling consists in sampling the system at a regular interval of a CV by adding a confining quadratic bias along this CV. 60 copies (called windows) of the system are used for that, each bias of each window has to be adjacent to neighbour ones. The standard CV used for this step is a s_{12} (i-e 12-referenced PCVs) defined from the CA data, the process to obtain it is detailed in the next subsection.

Once 15ps of trajectories has been sampled into each window (which represents 900ps of data), a systematic check is performed to avoid hidden errors, such as hysteresis. These errors take the form of separations in the coordination space between adjacent windows. In that case that indicates that a part of the chemical process is missed a part of the reaction is not sampled. Most of the time this separation appeared directly in the plot of the umbrella sampling data with s_{12} and z_{12} as coordinates. Yet in our verification process, we checked every pair of coordination that is used in the reference tables, with s_{12} and z_{12} as well, for each step. Two types of separation can be observed :

- ▷ When the gradient of the free energy is very important between two references, then a separation along s_{12}, visible in the data points, could occur. To correctly fill this hole into the dataset we insert new windows between the two separated ones with a stronger spring constant in order to counter the landscape gradient. But the strength of the quadratic bias you add into the system is limited by the timestep of 0.5fs.

- ▷ When some external degrees of freedom can evolve and change the position of an entire windows apart from one of its two adjacent windows in the phase space we call that an hysteresis. If no separation between adjacent windows is observed then the data-set is validated for quantification. In the other case then two possibilities can be distinguished :

  - The separation is on a degree of freedom that does not concern the reaction (i-e the $NH_3 \rightleftarrows [NH_4]^+$ long equilibrium in water). Then we can avoid it by forcing the

system to stay in the first state with an external potential wall. It is a situation better to avoid but that does not threaten the quality of the results if the few points that are modified by the external bias are taken out of the data set before quantification.

– The separation is on a degree of freedom that concerns the reaction. This means it is a normal evolution of the system toward relaxation and our s_{12} does not succeed into following. In that case, for the unique time it occurs, we decided to redefine s_{12} on a more reliable data-set using shooting from top, and relaunch a new US. We discuss more about this peculiar process in the next subsection.

Most of the time a projection of the dataset on s_{12} and z_{12} variables is sufficient to observe the undesirable hysteresis.

Shooting from top

Shooting from top is an algorithm that generates new trajectories of transition using a first existing one and an order parameter. We applied this algorithm for the last SN2 process of the synthesis. The algorithm :

1. Extract geometries from the actual transition trajectory that are inside a [A,B] interval of the order parameter
2. Pick a geometry in that set (uniform distribution)
3. Initiate random velocities with Boltzmann distribution
4. Launch 2 trajectories one forward, one backward (inverse velocities)
5. Wait it falls into reactants or products
6. Merge them
7. • IF a new transition is obtained and IF a metropolis criteria is respected[11] :

- THEN new transition became the actual transition
- ELSE go to 2.

Following this process, we generate 600 runs with a rate of success of 1/3. The [A,B] interval was [1.35,1.55] on s_2, including the supposed values for the transition state. We took out of this the last accepted trajectory and launched a new committor analysis from it.

Reference Space Exploration Algorithm

This protocol creates a reference list for $s_1 2$ and $z_1 2$ with two main constraints :

- Reference R_2 and R_{11} are fixed to correspond to the average among reactants and products ensembles. This implies that local minima in the free energy landscape are situated at values of s_{12} equal to 1 and 11. These values approximately correspond to $\approx$0.1 and $\approx$0.9, when rescaled.
- The list $[R_3 : R_{10}]$ of references between those to fixed pattern is optimised in order to obtain equivalent distances for each possible segment $[R_i, R_{i+1}]$, and equivalent angles for each possible set $\{R_i, R_{i+1}, R_{i+2}\}$

The two side references (respectively R_1 and R_{12}) are non-physical linear extensions of the pattern in the reactant and products ensembles in order to characterise the stable states (at R_2 and R_{11}) by true local minimums of the free energy curve. All technical details, about this algorithm and its usage, are detailed in our preceding paper.[8] Highly referenced PCVs has been used in US for steps with a level of complexity higher than a deprotonation. As deprotonation is straightforward, we preferred to use the proton coordination of the concerned atom as RC.

Machine Learning Interatomic Potential (MLIP) training

The aim of such techniques is to learn the potential energy surface from positions, energies and forces generated during *ab initio* simulations. It has been successfully used to improve

the study of equilibrium systems, but it is however a tremendously harder task to train a MLIP for reactive systems.[9,12,13] Indeed, the MLIP needs to be accurate on the whole configuration space of the chemical system, and therefore, the training set must be representative of the transition mechanism. Nonetheless, the training points must be carefully chosen in order to save computational time: if all the US simulations are placed in the training set, the MLIP will most certainly have good performance all along the RC-space, but it will have no use, since all the information we want is already included in the training set.

We train our MLIP into decomposing the energetic and forces associated with a position into atomic contributions of all the species in the bulk. In order to take into account the strong heterogeneity of the simulation box, we introduce the following loss:

$$L(\mathbf{w}) = \frac{1}{|B|} \sum_{l \in B} \left[p_E |E_l - E_l^{\mathbf{w}}|^2 + p_f \frac{1}{N_{elem}} \sum_{i=1}^{N} \frac{1}{n_i} |\mathbf{F}_{l,i} - \mathbf{F}_{l,i}^{\mathbf{w}}|^2 \right] \tag{5}$$

where n_i is the number of atoms of the same element as atom i in the system, N_{elem} is the number of different elements in the system, E_l and $F_{l,i}$ denote the DFT energies and forces of the training set, while $E_l^{\mathbf{w}}$ and $F_l^{\mathbf{w}}$ are the forces and energies computed by the MLIP, and B is the batch size (i.e., the number of geometries in the training set). $\mathbf{w}$ denotes the set of parameters of the neural networks. By weighting with n_i^{-1} the force-related terms, we ensure that each atomic species has the same weight in the training process. Using this loss, several neural network potential (NNPs) are trained using the DeePMD-kit smooth edition,[14,15] together, they form a committee. During a simulation, the maximum deviation on the prediction of the forces expressed in equation 6 is used as a metric to control the accuracy of the prediction during a simulation.

$$\sigma_{max}(t) = \max_{i \in [1, N_{atoms}]} \sqrt{\sum_{k=1}^{4} |\mathbf{F}_{i,t}^{(k)} - \overline{\mathbf{F}_{i,t}}|^2} \tag{6}$$

where 4 is the committee size, $\mathbf{F}_{i,t}^{(k)}$ is the force predicted for atom i by committee member k, $\overline{\mathbf{F}_{i,t}}$ is the average prediction of the committee.

Using this, a maximum simulation lifetime τ can be defined as the time σ_{max} stays below a given threshold. With this definition, simulations can be performed with the NNP committee and identify the zones of the RC-space where τ is too low. An AIMD US simulation is performed in the window at the center of these zones and put in the training set to train a new NNP. This is done until τ is the same in every US window.

Once convergence in terms of τ is achieved, we start the production runs simulations. In every window, from the same starting configuration as for the AIMD runs, we start NNP driven US simulations. In order, to control the behaviour of the system and to ensure the NNP stays within the region in which the forces are evaluated accurately, we perform the "mirror reflection" trick presented in ref.[9] By reflecting the velocities of the system on the surface where the deviation of the forces is the highest we are able to sample long stable trajectories.

Table S-2: Cost of one US step. Each line corresponds to a different perspective on the cost. The first is the time consumption on the calculation center, the second one is related to energetic consumption, and the last one to the environmental impact.

	AIMD	MLMD	Diff
Resources (kCPUh)	360	64	-296
Consumption* (kWh)	3600	640	-2960
CO_2^{**} (kg)	122	22	-100

* Consumption of the jean-zay CPU partition is of 10Wh.CPUh^{-1}.

** The average CO_2 per kWh in France in 2021 was 34g.kWh^{-1} according to RTE..[16]

Table S-2 provides a preliminary estimate of the cost for our calculations for a single FEP.

Full protocol flowchart

The figure S-3 represent our full protocol in a schematic way. A detailed technical discussion about this protocol is available in (See section **S-1**). The initial geometries of the reactants and products have been generated, optimized and equilibrated to the targeted temperature

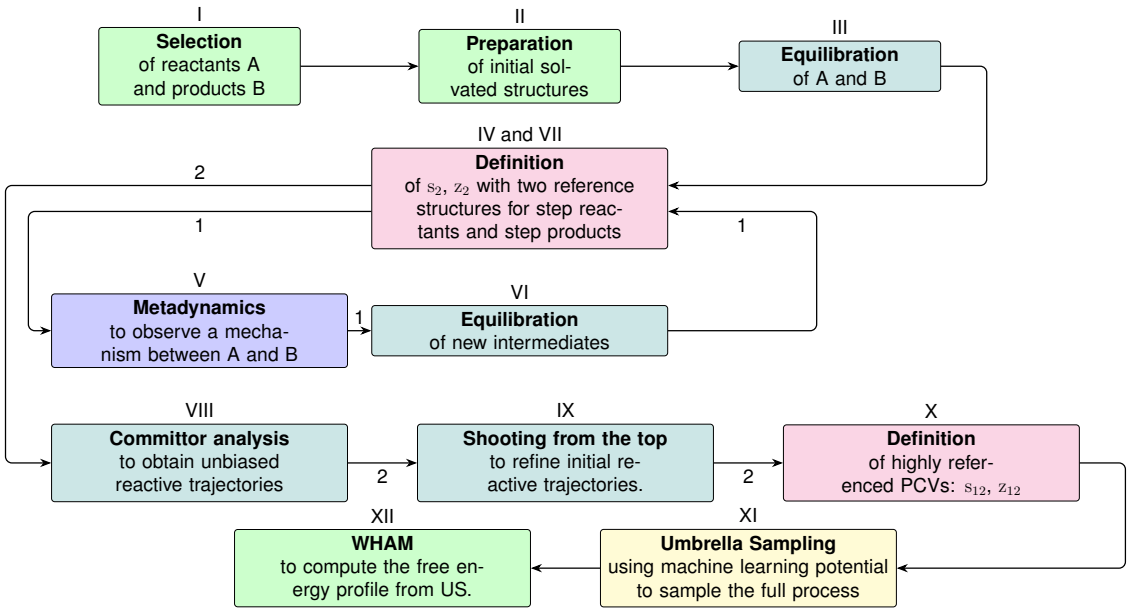


Figure S-3: Schematic algorithm depicting our simulation protocol. Green blocks (I, II, XII) indicate pre- and post-processing blocks of the protocol. Blue blocks (III, VI, VIII and IX) indicate *ab initio* agnostic explorative steps. Red blocks (IV/VII and X) indicate pivotal steps where new CVs are defined. The elementary acts that have been discovered during metadynamics (block V) pass one by one in the steps that follow: blocks VI to XII.

using the Gromacs software[17] with the AMBER force field[18] for solutes compounds, and the TIP3P model for water.[19]

Umbrella Sampling check process

Here we show a set of the couple of variables we plot after an Umbrella sampling process in order to check the quality of the results. For that we choose the (2') $\rightarrow$ (3) step. The plots are in figure S-4. This check process is the most time consuming of the protocol and could be accelerated or automatised by an algorithm in future works.

Free energy curves: WHAM algorithms

This method permits to obtain free energy curve out of independent molecular dynamic trajectories with static bias. Therefore it is adapted to umbrella sampling FES extraction. It has been performed separately with two different software in order to increase external codes independence of the results. The first one comes from Alan Grossfield,[20] and the second one from Andrew L. Ferguson.[21] To estimate the error due to convergence, which is always much higher than WHAM resolution error, each data set of 15ps of US is separated in four time intervals of 3.75ps each. The first two intervals are discarded for convergence. The difference between the two last obtained curves, divided by $\sqrt{2}$ according to block average method, gives an upper bound of the convergence error and the mean value of the two last curves gives the final FES. An example of Grossfield/Ferguson results is presented in figure S-5.

Non-agnostic check of the last transtition state

To test the counter-intuitive last transition state we decided to lock in in at the top of the barrier. To performed that we relied on a fully euristic CV: the difference of distance between

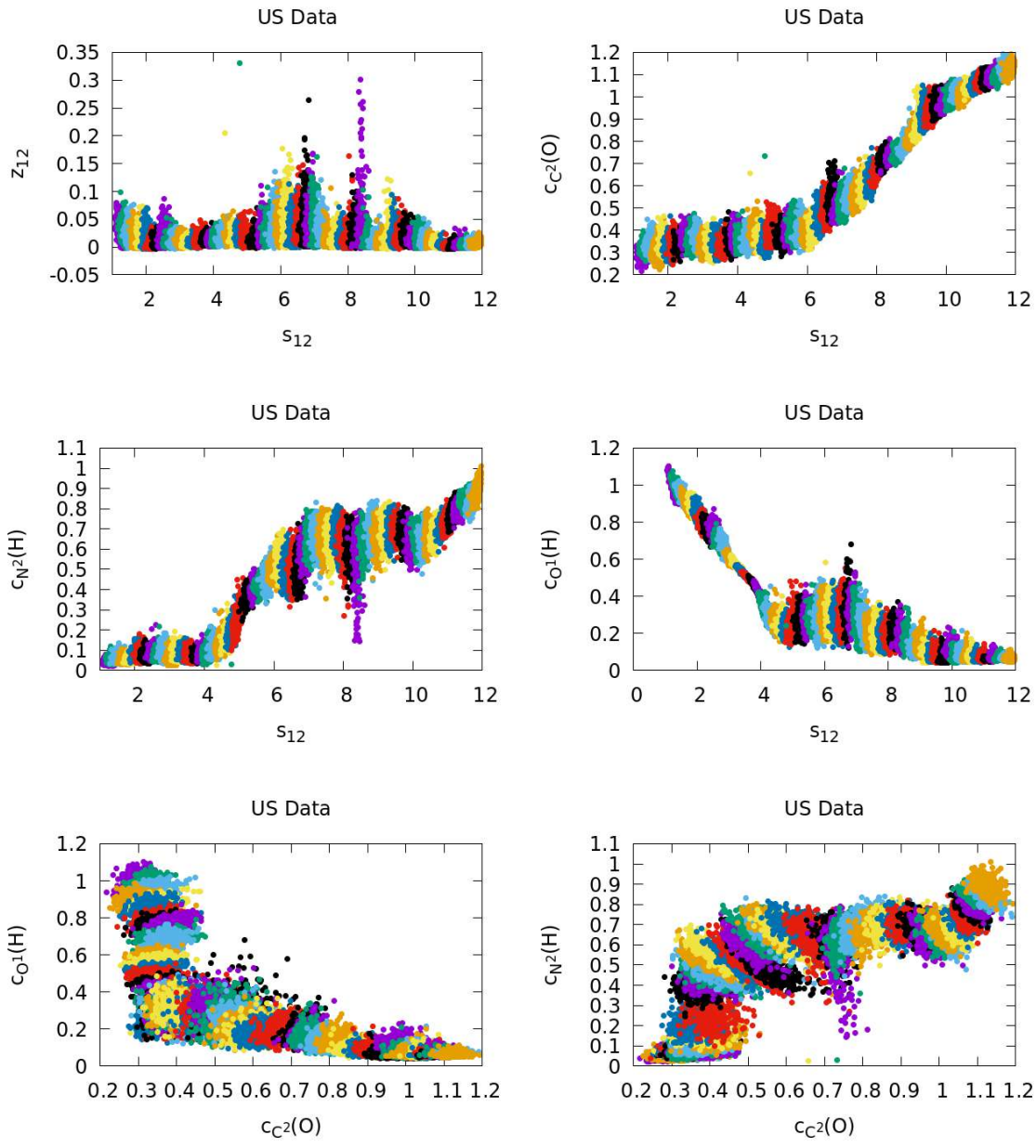


Figure S-4: Projections of the full (2') → (3) US Dataset for some couple of variables: s_{12}, z_{12}, $c_{C^2}(O)$, $c_{N^2}(H)$, $c_{O^1}(H)$. The full number of variable tested in our process is 28 for this step.

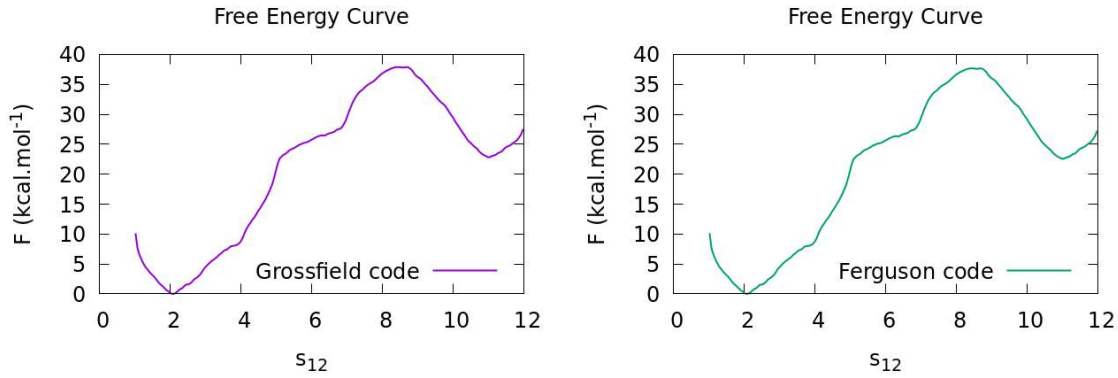


Figure S-5: Free energy curves for (2') → (3) reaction with Grossfield and Ferguson WHAM codes

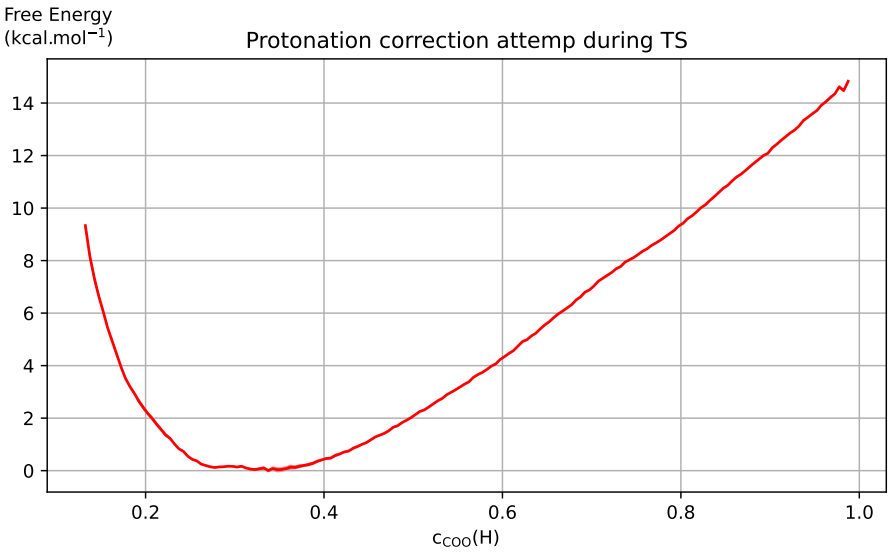


Figure S-6: The free energy curve we obtained by trying to sample protonated geometries of the last transition stated, locked in the reactive region by a quadratic bias. The study of the curve reveal that in our case there isn't any stable configuration of the protonated structure. The TS evolves spontaneously toward basic forme.

the nucleophile (nytrogen) and the living groupe (oxygen). This CV had been shown to be very efficient in the case of S_N2 like reactions. We locked this CV on the value of 0 with a harmonic bias with a strength of 0.4Ha (251 kcal.mol^{-1}.bohr^{-1}), the same order of magnitude of the bias we add during US.

As shown in figure S-6, the new US analysis we performed on the protonation of the acid group shows that there is no stable protonated form. This could be due to the size of the system, the protonation of the living hydroxide can be favoured by the releases of one proton in the solvent making passing the pH form 7 to ≈ 0.7 without intermediate values.

Hybride DFT study for dataset

The dataset we use for neural network potentials training are roughly of 0.1 ns lenght with a time step of 0.5fs. that represents 200k geometries. By calculating thees geometries via PBE and re-estimate the energy of a part of them thanks to PBE0 we can hope to be able to create a sufficient dataset of PBE0 energies and forces especially if we try to select wisely the recalculated geometries. That would allow use to train a "PBE0-trained" NNP and to obtain hybrid accuracy without any further calculations than the "few" PBE0 estimations. This section presente a first rought estimation of the cost of this training set generation for CPMD[22] and CP2K[23] softwares.

Table S-3: Time before SCF convergence for our 253 atoms system on a TS geometry.

xc	CPMD	CP2K
PBE	49	306
PBE0	1568	606
ratio	32	2

CP2K and CPMD sofwares do not use the same basis sets (Planewaves and Hybrides) and it is very hard to make them equivalent but we can still compare the time ratio of this two softwares when passing from PBExc to PBE0xc. The PBE SCF wavefunctions are initialise on atomic first guess. For PBE0, a first wavefunction optimisation with PBE (counted in the full time) is firstly performed to generate an initial wfc restart file.

As shown in table S-3, the usage of PBE0 with CPMD is highly non-recommended as

multiplying the calculation time by factor of 40, even with a wise wavefunction initialisation with PBE. On the contrary the usage of PBE0 with CP2K again with a "wise" wave functions initialisation only multiply it by 2. Stated on this preliminaries we recommend to used CP2K. The time of CP2K PBE0 calculation can still be improved by optimising it (exemple by chosoing an adapted basis set) so we can hope to reach a relatively affordable dataset if we succeed to divide its size by at least 10. This could be achieved by wisely choose the geometries for the training in a PBE generated dataset of geometries.

Table S-4: PBE0 relative error for randomly taken geometries of the reactants transition state and products of the last step toward glycine

	PBE	PBE0	%Err
$\Delta E^{\ddagger}$	+11.04	+18.25	40%
ΔEr	−14.38	−17.01	15%

In order to estimate what would this change made on the potential energy of our model we made, a short and rough estimation of the difference of behaviour of PBE and PBE0 on our system by selecting randomly one reactant geometry one transition state geometry and on product geometry of the last step calculate PBE and PBE0 energies for all of thees geometries and estimate the energy gap involved, see table S-4. This energy gaps are far from the free energy ones of the paper because they are not taking in account entropy effect nor solvent fluctuation effects as with AIMD. We measured ≈7 kcal.mol^{-1} of difference on the transition states energy and ≈3 kcal.mol^{-1} on reaction energy.

Protocol Step Parameters

AIMD

The parameters for unbiased trajectories are all in the input template in subsection **S-1-1**. For these calculations the PLUMED plugin is only used to monitor the values of several important CVs. For each exploration of each intermediate, the trajectories are 15ps length (≈30 000 steps).

Metadynamics

The gaussian bias sum in metadynamics is incremented every 50 steps ($\approx$ 25fs). Each gaussian 2D function is 0.005au height ($\approx$ 3.1kcal·mol^{-1}), 0.03 wide for s_2 and 0.2 wide for z_2. The metadynamics trajectory is 60ps length.

Committor Analysis

Committor analysis are run around probable transition region every 10 step of the initial trajectory. 3 launches of new trajectories for each geometries are initially performed. The geometries that do not commit in both reactant and products are ignored, the other are completed up to 20 launches. The wells are identified via s_2 and z_2 PCVs specifically parameterised for each step. The Committor Analysis trajectories are roughly of 1.6ps length each which is voluntarily much higher than the normal commit time ($\approx$ 100fs).

Shooting From the Top

The interval to launch Shooting From Top has been chosen to be the 1.35 to 1.5 region of the corresponding s_2 for the last and unique step for which it has been used: the $6 \rightarrow 7$ one. We chose this interval as the one of possible commit toward both wells observed in the preceding Committor Analysis. We launched 600 successive transition attempts both forward and backward. The acceptance ratio was 1/3, which means that 200 new probable transition state geometries have been identified during the process, each one more independent from the original metadynamic one.

Umbrella Sampling

For complex transitions, 60 umbrella windows are launched on the s_{12} corresponding with the transition, with a strength constant k = 0.18au ($\approx$110kcal·mol^{-1}). The bias is then calculated using equation 7, were s0 designate the center of each window. They are equally

spread into the 1 to 12 interval of the s_{12}.

$$Bias(s_{12}) = k/2(s_{12} - s0)^2 \tag{7}$$

For deprotonation steps, we launched 15 umbrella windows on the proton coordination of the deprotonated atom, with a strength constant k = 1.8au (≈1000kcal·mol^{-1}). Windows are equally spread from the 0 to 1 values of the proton coordination.

Supporting References

References

(1) Paesani, F. Hydrogen bond dynamics in heavy water studied with quantum dynamical simulations. *Physical Chemistry Chemical Physics* **2011**, *13*, 19865.

(2) Zheng, H.; Wang, S.; Zhang, Y. Increasing the Time Step with Mass Scaling in Born-Oppenheimer ab initio QM/MM Molecular Dynamics Simulations. *Journal of computational chemistry* **2009**, *30*, 2706–2711.

(3) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *The Journal of chemical physics* **2007**, *126*, 054103.

(4) Magrino, T.; Pietrucci, F.; Saitta, A. M. Step by Step Strecker Amino Acid Synthesis from Ab Initio Prebiotic Chemistry. *The Journal of Physical Chemistry Letters* **2021**, *12*, 2630–2637, Publisher: American Chemical Society.

(5) Pietrucci, F.; Saitta, A. M. Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios. *Proceedings of the National Academy of Sciences* **2015**, *112*, 15030–15035, Publisher: Proceedings of the National Academy of Sciences.

(6) Brigiano, F. S.; Gierada, M.; Tielens, F.; Pietrucci, F. Mechanism and Free-Energy Landscape of Peptide Bond Formation at the Silica–Water Interface. *ACS Catalysis* **2022**, *12*, 2821–2830, Publisher: American Chemical Society.

(7) Saitta, A. M.; Saija, F. Miller experiments in atomistic computer simulations. *Proceedings of the National Academy of Sciences* **2014**, *111*, 13768–13773.

(8) Magrino, T.; Huet, L.; Saitta, A. M.; Pietrucci, F. Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry. *The Journal of Physical Chemistry A* **2022**, *126*, 8887–8900, Publisher: American Chemical Society.

(9) Devergne, T.; Magrino, T.; Pietrucci, F.; Saitta, A. M. Combining Machine Learning Approaches and Accurate Ab Initio Enhanced Sampling Methods for Prebiotic Chemical Reactions in Solution. *Journal of Chemical Theory and Computation* **2022**, *18*, 5410–5421, Publisher: American Chemical Society.

(10) Pérez-Villa, A.; Pietrucci, F.; Saitta, A. M. Prebiotic chemistry and origins of life research with atomistic computer simulations. *Physics of Life Reviews* **2020**, *34-35*, 105–135.

(11) Jung, H.; Okazaki, K.-i.; Hummer, G. Transition path sampling of rare events by shooting from the top. *The Journal of chemical physics* **2017**, *147*, 152716.

(12) Young, T. A.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chem. Sci.* **2021**, *12*, 10944–10955, Publisher: The Royal Society of Chemistry.

(13) Yang, M.; Bonati, L.; Polino, D.; Parrinello, M. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catalysis Today* **2022**, *387*, 143–149, 100 years of CASALE SA: a scientific perspective on catalytic processes.

(14) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications* **2018**, *228*, 178–184.

(15) Zhang, L.; Han, J.; Wang, H.; Saidi, W. A.; Car, R.; E, W. End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems. *arXiv:1805.09003 [cond-mat, physics:physics]* **2018**, *31*, arXiv: 1805.09003.

(16) RTE, BilanElectrique2021 _ Synthese. 2021; `https://bilan-electrique-2021.rte-france.com/synthese-les-faits-marquants-de-2021/`.

(17) Abraham, M. *et al.* GROMACS 2023.1 Source code. 2023; `https://zenodo.org/record/7852175`.

(18) Kollman, P. Amber 2023. 2023.

(19) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(20) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.11. `http://membrane.urmc.rochester.edu/?page_id=126`, Last accessed 28 September 2022.

(21) Ferguson, A. L. BayesWHAM: A Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. *Journal of Computational Chemistry* **2017**, *38*, 1583–1605, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.24800.

(22) Hutter, J.; Curioni, A. Car–Parrinello Molecular Dynamics on Massively Parallel Computers. *ChemPhysChem* **2005**, *6*, 1788–1793, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.200500059.

(23) Kühne, T. D. *et al.* CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **2020**, *152*, 194103.

Appendix D

Correction to "Step by Step Strecker Amino Acid Synthesis from ab Initio Prebiotic Chemistry"

Correction to "Step by Step Strecker Amino Acid Synthesis from ab Initio Prebiotic Chemistry"

Léon Huet,* Théo Magrino, Fabio Pietrucci, and A. Marco Saitta

Sorbonne Université, Muséum National d'Histoire Naturelle, UMR CNRS 7590, IMPMC, Paris, France

E-mail: leon.huet@sorbonne-universite.fr

Correction to "Step by Step Strecker Amino Acid Synthesis from ab Initio Prebiotic Chemistry", Théo Magrino, Fabio Pietrucci and A. Marco Saitta
J. Phys. Chem. Lett. 2021, 12, 10, 2630–2637
DOI : 10.1021/acs.jpclett.1c00194

Van der Waals empirical correction

In the cited work,[1] Grimme's van der Waals (vdW) corrections[2] were accidentally overlooked and not included in the potential calculations. Our goal is to address this oversight to prevent its recurrence in future studies. The oversight was originated from the vdW section of the CPMD software input file, with no warning in version 4.1 of the CPMD software.[3]

```
Former version :
...
& VDW
  VDW PARAMETERS
  ALL DFT-D2
  S6GRIM
  PBE
&END
...
Corrected Version :
...
& VDW
 EMPIRICAL CORRECTION
  VDW PARAMETERS
  ALL DFT-D2
  S6GRIM
  PBE
 END EMPIRICAL CORRECTION
&END
...
```

The keywords "EMPIRICAL CORRECTION" and "END EMPIRICAL CORRECTION" are missing in the description of the vdW section of the CPMD manual.

To evaluate how this oversight initially impacted the article's conclusions, we performed a new umbrella sampling (US) simulation for the transition from step **(3)** to **(4)** using the corrected input file. We chose this step for two reasons: it depends on a long-range interaction between two reactive molecules that could be influenced by the omission of vdW corrections, and it has been used as a reliable benchmark reaction for additional simulations.[4]

The protocol for the revised **(3)**→**(4)** US simulation remained the same as in the reference:[1] in each window we generated 14 ps of trajectory, divided into 7 ps for equilibration and 7 ps for free-energy sampling. The results, shown in figure E.1, indicate consistency between the two calculations, with $\Delta_r F = -23 \pm 0.5$ kcal.mol^{-1} for both. As anticipated for condensed phase chemical reactions, empirical vdW corrections have minimal impact on the results. Additionally, we confirmed that the corrected input file accurately implements Grimme's van der Waals D2 corrections and that these are properly included in the force estimates. We expect that incorporating vdW corrections in other steps of the mechanism will still yield results within the error margins of the initial study.[1]

Van der Waals corrections are consistently accounted for also in the simulations described in the following sections.

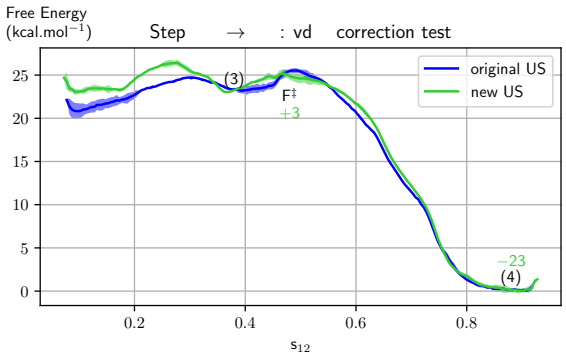


Figure E.1: Results of the new US simulation with Grimme's D2 correction. The new free-energy curve is depicted in green, while the the original one, as presented in the article, is presented in blue. The reactant position for the step $s_{(3)\rightarrow(4)}$ is located at 0.4 on the s_{12} axis. The error bars are indicated for both with shaded areas.

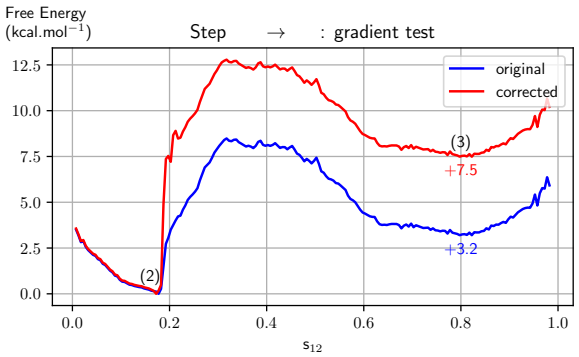


Figure E.2: Effect of the addition of an US window on the **(2)→(3)** step. The blue curve, as presented in the article, represents the original data, while the red curve includes the effect of an additional window.

Correction concerning a strong free-energy gradient in the (2)→(3) step

An issue was identified during the sampling of the **(2)→(3)** step in the Strecker mechanism, as described in ref,[1] which involves the formation of a cation through the elimination of a water molecule. A pronounced gradient was noted in a specific region of the free-energy curve. After incorporating additional US windows in this region, the free-energy landscape was impacted, increasing the final free-energy difference between reactants and products by 4 kcal.mol^{-1}, as shown in figure E.2. The s_{12} coordinate initially designed for this step seems inadequate to capture this chemical transformation, thus requiring a further determination of the free-energy curve. Consequently, we defined a new set of coordinates, s_{12} and z_{12}, based on new committor analysis and proceeded to repeat the entire US simulation for this step to verify the accuracy of the results.

Corrections to hysteresis effects in the (1)→(2'), (2)→(3), (4)→(5') and (5')→(5) steps

After carrying out a detailed analysis of the US data, we spotted hysteresis effects that were initially overlooked, as shown in figure E.3. These effects were evidenced through the determination of new collective variables and a continuity check, as described in our protocol paper.[5] For each step, we established new pairs of s_{12} and z_{12} variables and repeated the US simulations. Notably, Step **(1)→(2')** was subdivided into two distinct processes: the HCN deprotonation **(1)→(1')**, and the amine addition **(1')→(2')**. The deprotonation reaction coordinate was defined by the coordination of the carbon with hydrogen, which underlines the relative simplicity of the reaction.

The new, revised free-energy curves for this process are reported in figure E.4. The updated diagram is presented in figure E.5. We have not identified additional errors of this nature in the other steps.

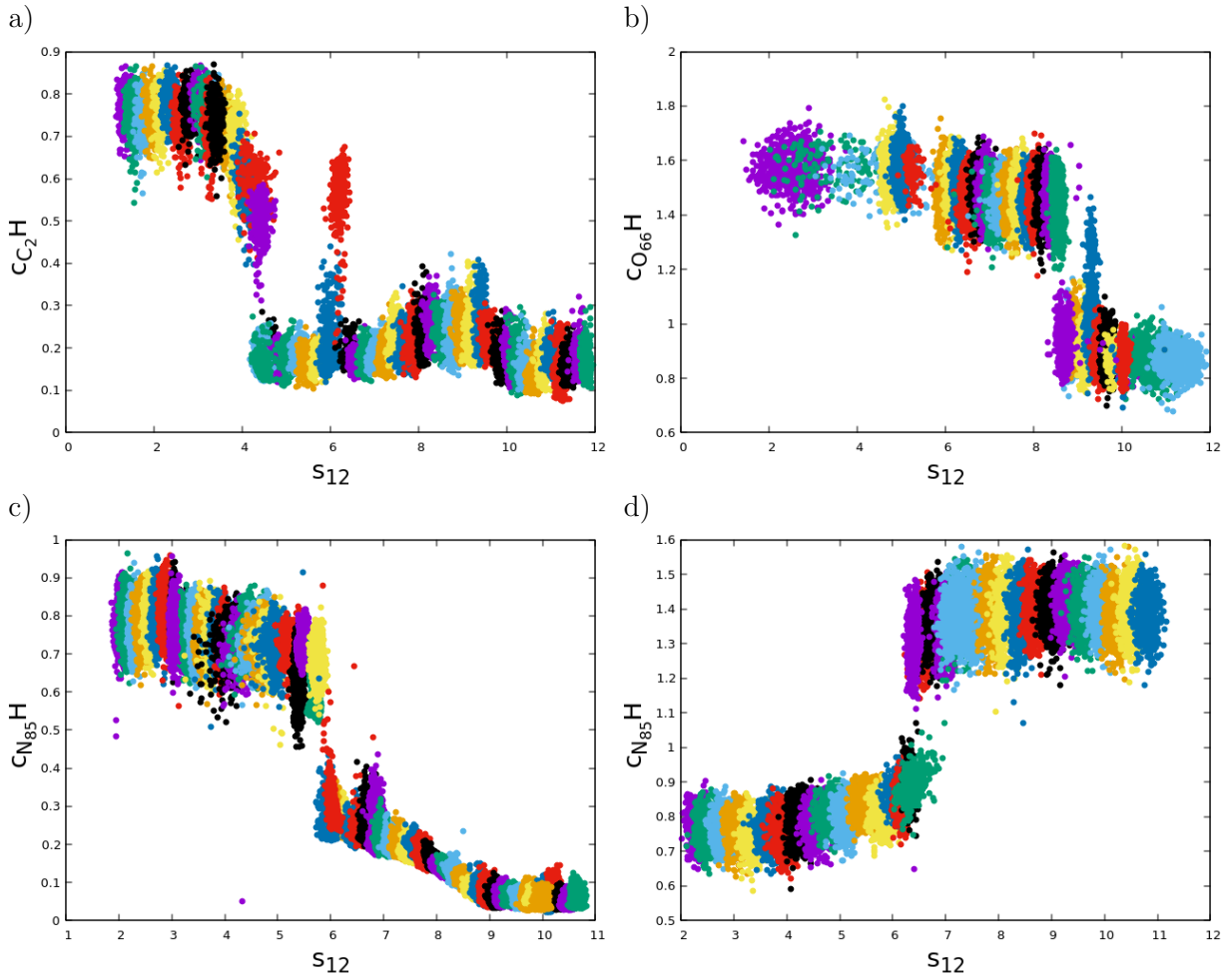


Figure E.3: Hysteresis has been observed in previous US simulations at several steps: **a)** In step **(1)→(2')**, hysteresis was detected in the protonation of the cyanide's carbon, $c_{C_2}(H)$. **b)** In step **(2)→(3)**, the critical variable is the protonation of the leaving oxygen, $c_{O_{66}}(H)$. **c)** For step **(4)→(5')**, the critical variable involves the protonation of the leaving nitrogen, $c_{N_{85}}(H)$. **d)** In step **(5')→(5)**, the critical variable is again the protonation of the leaving nitrogen, $c_{N_{85}}(H)$. The s_{12} coordinate is used as the horizontal axis for each graph.

Bibliography

References

(1) Magrino, T.; Pietrucci, F.; Saitta, A. M. Step by Step Strecker Amino Acid Synthesis from Ab Initio Prebiotic Chemistry. *The Journal of Physical Chemistry Letters* **2021**, *12*, 2630–2637, Publisher: American Chemical Society.

(2) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006**, *27*, 1787–1799.

(3) Hutter, J.; Curioni, A. Car–Parrinello Molecular Dynamics on Massively Parallel Computers. *ChemPhysChem* **2005**, *6*, 1788–1793.

(4) Devergne, T.; Magrino, T.; Pietrucci, F.; Saitta, A. M. Combining Machine Learning Approaches and Accurate Ab Initio Enhanced Sampling Methods for Prebiotic Chemical Reactions in Solution. *Journal of*

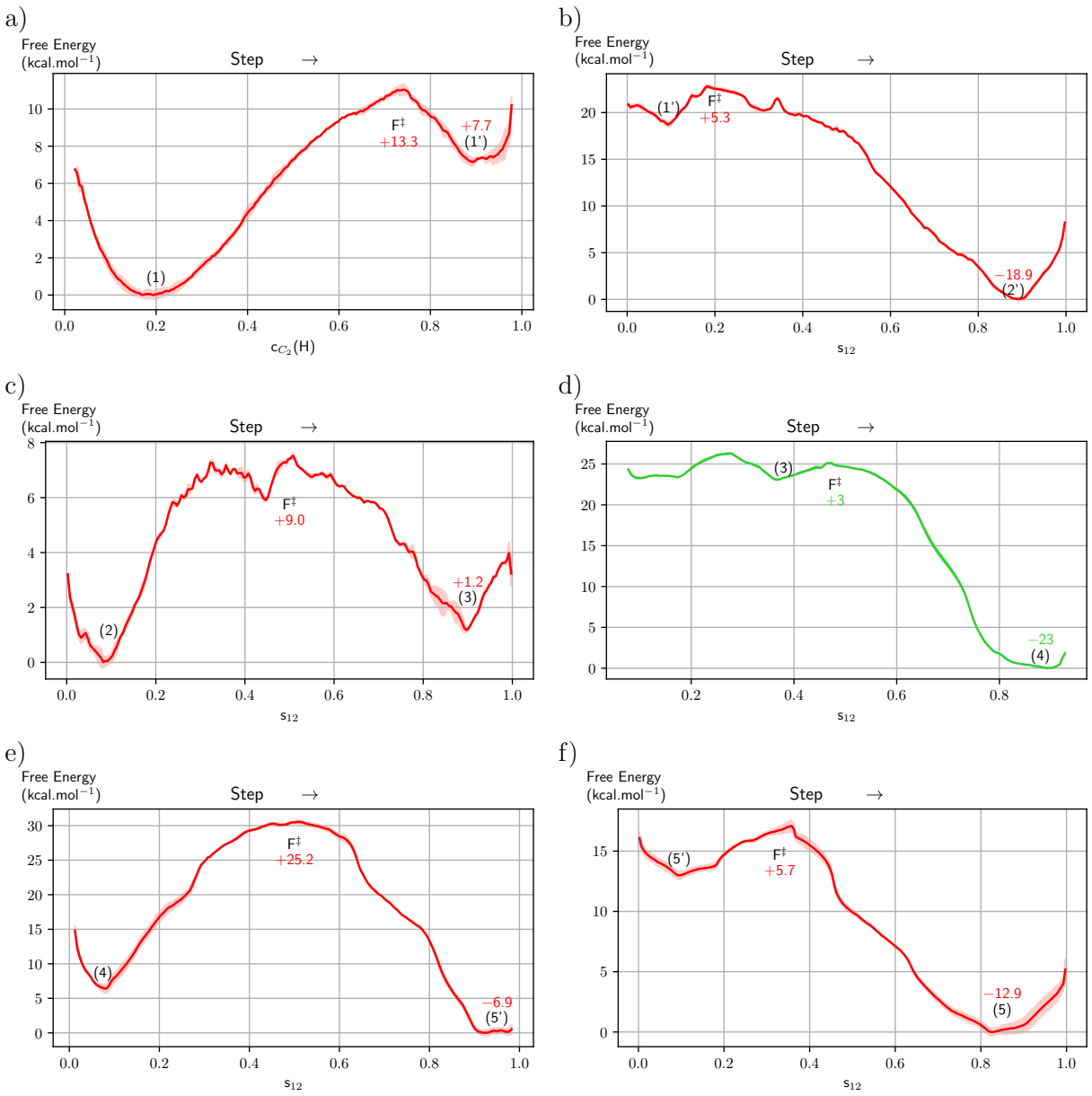


Figure E.4: Free-energy curves calculated out of the newly performed US simulations, with error bars. The one made to check vdW correction effects is represented in green. The ones performed to correct the observed hysteresis in the original data are drawn in red.

Chemical Theory and Computation **2022**, *18*, 5410–5421, Publisher: American Chemical Society.

(5) Magrino, T.; Huet, L.; Saitta, A. M.; Pietrucci, F. Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry. *The Journal of Physical Chemistry A* **2022**, *126*, 8887–8900, Publisher: American Chemical Society.

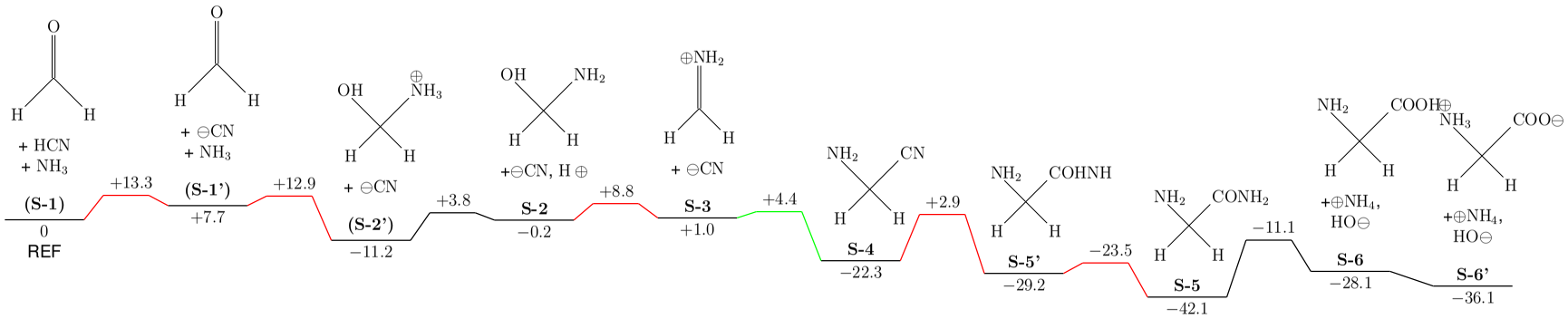


Figure E.5: New reaction diagram after the correction of the free-energy curves. The color code is consistent with figure E.4: in green the section remade to check effect of vdW corrections omission in the original paper,[1] in red the sections that regard hysteresis corrections for which we had to redefine new s_{12} and z_{12} coordinates.

Appendix E

Supporting information of "Insight on chemical reaction dynamics and reaction coordinates from non-Markovian models"

Supplementary Information for: Insight on chemical reaction dynamics and reaction coordinates from non-Markovian models

Léon Huet,[†] Hadrien Vroylandt,[‡] Rodolphe Vuilleumier,[¶] A. Marco Saitta,[†] and Fabio Pietrucci[*,†]

†*Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS UMR 7590, Paris 75005 FRANCE*

‡*Institut des sciences du calcul et des données, Sorbonne Université, 75005 Paris, France*

¶*PASTEUR, Département de Chimie, Ecole Normale Supérieur, PSL University, Sorbonne Université, CNRS, 75005 Paris, France*

E-mail: fabio.pietrucci@sorbonne-universite.fr

S-1 CPMD input template

```
&CPMD
  MOLECULAR DYNAMICS BO
  CONVERGENCE ORBITALS
   1.e-5
  RESTART GEOFILE VELOCITIES
  VDW CORRECTION
  STORE WAVEFUNCTIONS
   100
  EXTRAPOLATE WFN
   4
  TIMESTEP
   20.
  TEMPERATURE
   300.
```

```
  NOSE IONS MASSIVE
  300. 3000.
   TRAJECTORY XYZ SAMPLE FORCES
   1
  stress tensor
   20
  MAXSTEP
   1500000
  MAXRUNTIME
   3000
  COMPRESS WRITE32
  MIRROR
  RESTFILE
   2
  MEMORY BIG
  REAL SPACE WFN KEEP
  INITIALIZE WAVEFUNCTION RANDOM
  ALLTOALL SINGLE
  CP_GROUPS
   1
  PRNGSEED
   XXX
&END

&DFT
  FUNCTIONAL PBE
  GC-CUTOFF
   1.E-05
&END
&VDW
 EMPIRICAL CORRECTION
 VDW PARAMETERS
 ALL DFT-D2
 S6GRIMME
 PBE
 END EMPIRICAL CORRECTION
&END
&SYSTEM
  ANGSTROM
  SYMMETRY
   1
  CELL
   14.4807   1 1 0  0  0
  CUTOFF
   80
&END
&ATOMS
*C_MT_PBE  KLEINMAN-BYLANDER
 LMAX=P
CCC
*Cl_MT_PBE.psp  KLEINMAN-BYLANDER
LMAX=D
ClClCl
*H_MT_PBE  KLEINMAN-BYLANDER
```

```
LMAX=S
HHH
*K_MT_PBE_SEMI.psp  KLEINMAN-BYLANDER
LMAX=D
KKK
*O_MT_PBE  KLEINMAN-BYLANDER
LMAX=P
OOO
 ISOTOPES
  12.0107
  35.9768
   2.0
  38.9637
  15.9994
&END
```

S-2 Mass in the well without wall on cos(θ)

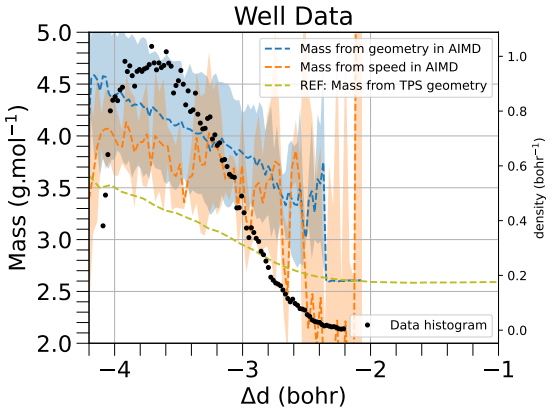


Figure S-1: Estimation of the mass of the variable in the reactant well in the absence of all on cos(*theta*)

The Figure S-1 shows the behavior of the mass of the variable in the absence of wall on the cosinus. We can observe that the error bars are still important after 100ps of unbiased sampling, which justify the usage of an additional wall to limit the exploration.

S-3 Reference frames of $s_1 2$ and $z_1 2$

The following lines are presenting the 12 references use to defined our data-driven variables. They are taken out of our preceding work on this system.[1] Lines are representing the atoms C, Cl_1 and Cl_2 and columns are representing the elements C, Cl, O,K and H. Each numerical value represents the coordinance of one of the three atoms in one of five elements.

```
NAT 3 FRAMES 12
1
  1  1      0.0000   0.8141   0.1336   0.0011   2.9176
  2  2      0.0829   0.0050   0.7436   0.0010   1.4340
  3  2      0.7312   0.0050   0.5988   0.0007   1.0591
2
  1  1      0.0000   0.8905   0.1327   0.0011   2.9233
  2  2      0.0906   0.0061   0.7274   0.0010   1.4188
  3  2      0.7999   0.0061   0.6015   0.0007   1.2040
3
  1  1      0.0000   0.9602   0.1572   0.0009   2.9254
  2  2      0.1634   0.0092   0.6924   0.0007   1.4212
  3  2      0.7968   0.0092   0.6404   0.0008   1.3503
4
  1  1      0.0000   1.0235   0.1667   0.0009   2.9751
  2  2      0.2069   0.0112   0.6426   0.0009   1.3007
  3  2      0.8166   0.0112   0.7178   0.0006   1.4023
5
  1  1      0.0000   1.0156   0.1486   0.0008   2.9459
  2  2      0.2665   0.0116   0.5688   0.0007   1.2175
  3  2      0.7491   0.0116   0.6954   0.0006   1.5018
6
  1  1      0.0000   0.9786   0.1818   0.0012   2.9417
  2  2      0.3596   0.0115   0.5981   0.0010   1.1912
  3  2      0.6190   0.0115   0.7144   0.0007   1.4740
7
  1  1      0.0000   1.0089   0.1594   0.0009   2.9360
  2  2      0.4891   0.0124   0.5882   0.0009   1.2141
  3  2      0.5198   0.0124   0.6729   0.0006   1.4508
8
  1  1      0.0000   0.9355   0.1503   0.0010   2.9715
  2  2      0.5232   0.0109   0.5788   0.0010   1.1887
  3  2      0.4123   0.0109   0.7059   0.0006   1.5428
9
  1  1      0.0000   0.8236   0.1515   0.0009   2.9703
  2  2      0.5295   0.0088   0.5792   0.0009   1.2277
  3  2      0.2941   0.0088   0.7454   0.0005   1.5164
10
  1  1      0.0000   0.8100   0.1520   0.0009   2.9887
  2  2      0.6426   0.0070   0.6134   0.0008   1.2507
  3  2      0.1675   0.0070   0.7593   0.0005   1.4876
11
  1  1      0.0000   0.9121   0.1496   0.0006   2.9362
  2  2      0.7758   0.0083   0.5969   0.0008   1.2332
```

```
  3  2      0.1362   0.0083   0.7620   0.0005   1.5268
12
  1  1      0.0000   1.0056   0.1502   0.0006   2.9401
  2  2      0.8758   0.0094   0.5940   0.0008   1.3485
  3  2      0.1298   0.0094   0.7613   0.0005   1.5316
```

S-4 Results of the study for additional variables

As 6 CVs was a too short sample to perform correlations on them we decided to add most of the CVs used to reference the $s_1 2$ and $z_1 2$. Some CVs that are links with Δd had also been added into the set: d_2 the other C-Cl distance and c_1-c_2 the coordinance equivalent of Δd.

Table S-1: Dynamics and quality features for supplementary CVs with varying qualities. The measured criteria are the same as for the 6 initial ones. The SLLS quantifies how likely a variable is to be linearly transformed into the committor at the top of the barrier. The transmission coefficient (TC) measures the recrossing effect during the transition. The subsequent parameters are derived from stochastic models. The friction coefficient and memory decay times are known to be inversely proportional to the quality of the variable to mimic the behavior of the committor. Error bars are estimated using block averages.

Var	SLLS	TC (%)	γ_1 (ps^{-1})	R_1^2 (-)	γ_2 (ps^{-1})	$t_{1/2}$ (fs)	γ_3 (ps^{-1})
c_1-c_2	17.2 ± 2.9	54.7 ± 3.3	51.6 ± 0.6	0.34 ± 0.01	22.7 ± 11.4	49.6 ± 7.5	10.0 ± 0.05
d_2	15.7± 5.2	48.3 ± 4.2	102.5 ± 4.0	0.30 ± 0.0001	13.1 ± 2.7	36.8 ± 6.3	17.2 ± 0.8
c_C(Cl)	0.0 ± 0.0	0.0 ± 8.9	53.6 ± 0.1	0.246 ± 0.006	23.7 ± 14.2	31.0 ± 1.9	11.4 ± 0.06
c_C(O)	1.9 ± 0.4	0.0 ± 7.6	88.0 ± 2.4	0.37 ± 0.038	44.1 ± 25.8	27.3 ± 2.2	16.3 ± 0.9
$c_{Cl_1}(Cl_2)$	1.16 ± 1.2	0.90 ± 0.2	47.8 ± 2.0	0.50 ± 0.05	23.5 ± 9.0	17.7 ± 5.1	5.1 ± 0.1
c_{Cl_1}(O)	0. ± 0.4	9.2 ± 5.1	64.2 ± 1.4	0.46 ± 0.02	32.1 ± 15.3	30.2 ± 5.1	11.5 ± 0.01
c_{Cl_1}(K)	0.2 ± 1.0	6.1 ± 6.2	27.2 ± 1.3	0.4 ± 0.1	150.1 ± 29.3	125.3±117.6	3.8 ± 0.2
c_{Cl_2}(O)	0.0 ± 0.8	9.3 ± 1.0	71.7 ± 0.5	0.23 ± 0.02	38.9 ± 0.3	29.3 ± 0.2	16.2 ± 0.3
c_{Cl_2}(K)	0.8 ± 0.4	16.3 ± 1.5	38.3 ± 5.1	0.48 ± 0.03	9.2 ± 8.6	354.8±331.1	4.4 ± 0.3

S-5 modification of the Jan Daldrop code

To add the effect of temperature in the Jan Daldrop code[2] in a intuitive way we modified the code directly in the python script we used. For transparency the new class we used are presented here:

```
import numpy as np
import bgle
```

```
from scipy.interpolate import CubicSpline

class ColoredNoiseGenerator_TEMP (bgle.ColoredNoiseGenerator):
    def generate(self, size, kBT):
        white_noise = self.rng(size=size)
        colored_noise = np.convolve(
            white_noise, self.sqk,
            mode='same') * np.sqrt(self.t[1] - self.t[0])
        return colored_noise * np.sqrt(kBT)

class BGLEIntegrator_TEMP(bgle.BGLEIntegrator):
    def __init__(self,
                 kernel,
                 t,
                 m=1.,
                 dU=lambda x: 0.,
                 add_zeros=0,
                 verbose=True):
        self.kernel = kernel
        self.t = t
        self.m = m
        self.dt = self.t[1] - self.t[0]
        self.verbose = verbose
        self.dU = dU

        if self.verbose:
            print("Found dt =", self.dt)

        self.noise_generator = ColoredNoiseGenerator_TEMP(
            self.kernel, self.t, add_zeros=add_zeros)

    def integrate(self,
                  n_steps,
                  kBT,
                  x0=0.,
                  v0=0.,
                  set_noise_to_zero=False,
                  _custom_noise_array=None,
                  _predef_x=None,
                  _predef_v=None,
                  _n_0=0):
        if set_noise_to_zero:
            noise = np.zeros(n_steps)
        else:
            if _custom_noise_array is None:
                noise = self.noise_generator.generate(n_steps, kBT)
            else:
                assert (len(_custom_noise_array) == n_steps) # type: ignore

        x, v = x0, v0

        if _predef_v is None:
```

```
        self.v_trj = np.zeros(n_steps)
    else:
        assert (len(_predef_v) == n_steps) # type: ignore
        assert (_predef_v[_n_0 - 1] == v)
        self.v_trj = _predef_v

    if _predef_x is None:
        self.x_trj = np.zeros(n_steps)
    else:
        assert (len(_predef_x) == n_steps) # type: ignore
        assert (_predef_x[_n_0 - 1] == x)
        self.x_trj = _predef_x

    self.t_trj = np.arange(0., n_steps * self.dt, self.dt)

    rmi = 0.
    for ind in range(_n_0, n_steps): # type: ignore
        last_rmi = rmi
        if ind > 1:
            rmi = self.mem_int_red(self.v_trj[:ind])
            last_v = self.v_trj[ind - 1]
        else:
            rmi = 0.
            last_rmi = 0.
            last_v = 0.

        x, v = self.rk_step(x, v, rmi, noise[ind], last_v, last_rmi)
        self.v_trj[ind] = v
        self.x_trj[ind] = x

    return self.x_trj, self.v_trj, self.t_trj
```

The difficulty is then to generate a trajectory with coherent units for the forces, the friction, the time step, the mass, and $k_B T$.

S-6 Reproduction of the well data

In order to extrapolate our GLE model, we first needed to test whether it could efficiently reproduce the dynamics of the variable, at least within the well ensemble. For this purpose, we performed sampling of the GLE model in the well with different time steps and temperatures. The results of this study are presented in Figure S-2. This study shows that the use of a linear rescale of the time step with respect to the temperature was sufficient to ensure correct integration at high temperatures. Theoretically, a rescale by $\frac{1}{\sqrt{T}}$ should have been

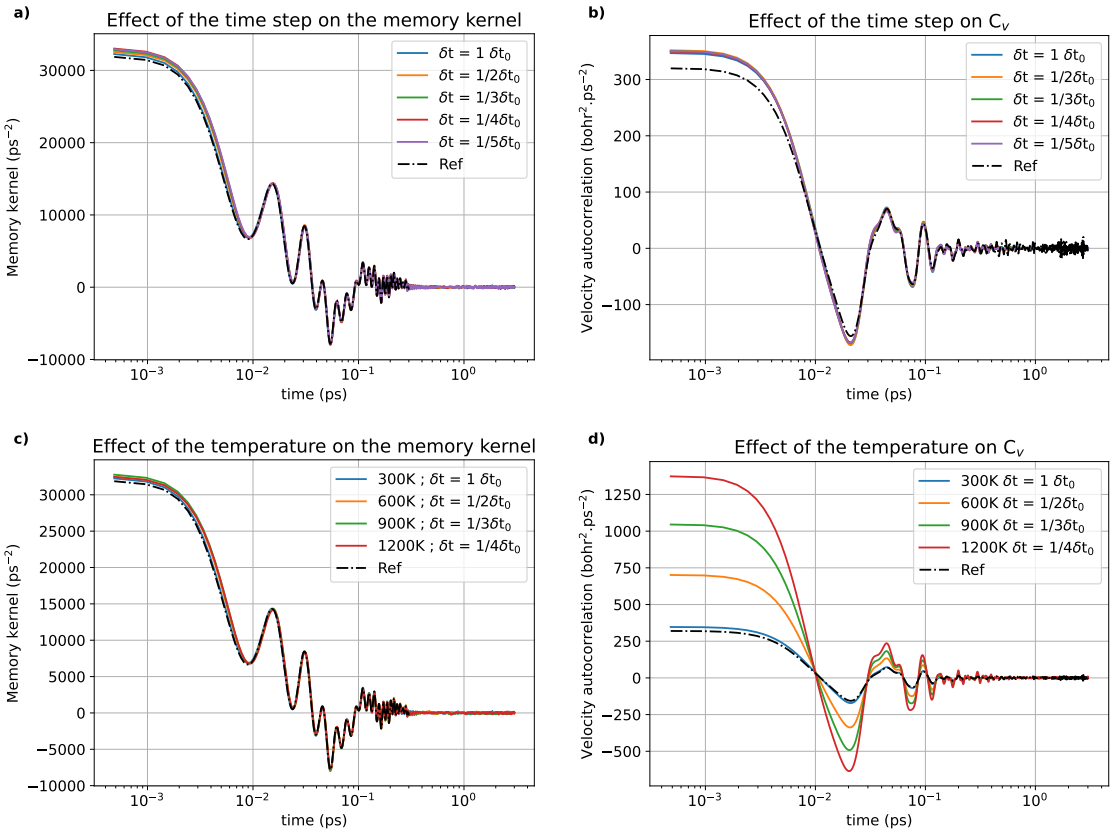


Figure S-2: Study of the effect of time step and temperature on the observed memory kernels and speed auto-correlation in the well ensemble for our GLE integration program. **a)** The effect of the sampling time step on the memory at 300K. A small decay is observable at short timescales and appears to be a regular error that the time step does not change. **b)** The effect of the sampling time step on the velocity autocorrelation. The observation is the same, showing a decay at short timescales that is unaffected by the time step. This could be due to the small mass difference between the top of the barrier and the well. **c)** The effect of temperature on the memory kernel. The friction is not affected. The use of a linearly rescaled time step seems to ensure correct sampling of the noise. **d)** The effect of temperature on the autocorrelation of the velocity. As expected, the autocorrelation intensity of the velocity evolves linearly with temperature. The sampling presents no artifacts with a linear rescale of the time step.

sufficient, but we decided to continue with the linear rescale as it was effective affordable.

S-7 MCCI datasets

As short representation of the TPS and US datasets is proposed from the point of view of Δd. The resulting graphs are presented in Figure S-3 and S-4.

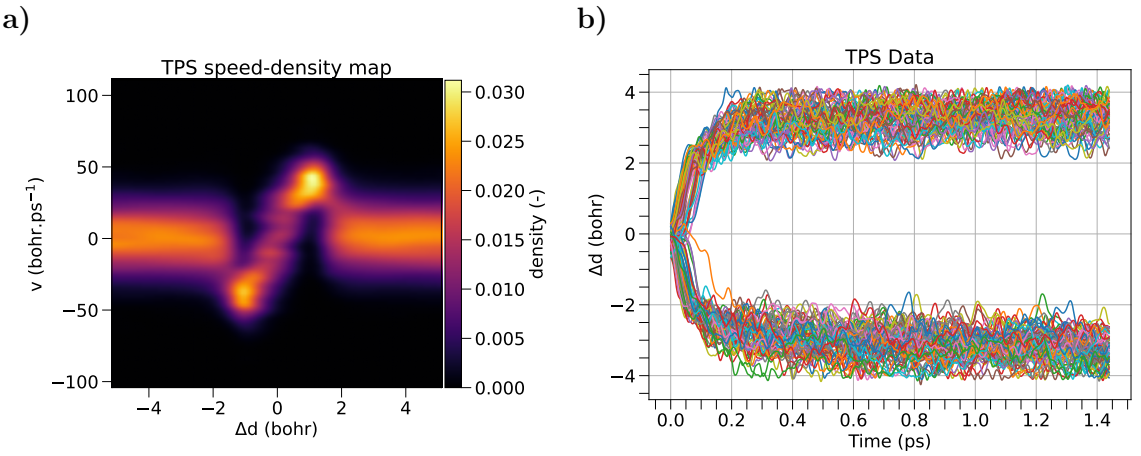


Figure S-3: Representation of the TPS dataset for the MCCI system. **a)** Distribution of the velocity conditioned on the value of Delta-d. We can see the "warm-up" of the variable due to the gradient descent from the top of the barrier (at Delta-d = 0). This behavior opposed to the Maxwell-Boltzmann distribution of the velocities is a characteristic of transition path sampling datasets. Outside of the [-2,2] interval in bohr, the velocity distribution is thermalysed, this means that we are in the well region. **b)** Time evolution of the shootings from the top trajectory. We can observe that the descent from the top takes roughly 0.2ps.

S-8 Effect of the wall on cos(θ) on the friction

We estimated the effect of the addition of the cos(θ) on the friction in the well data. We also compared γ_0 and γ_1. The results of this study is that the addition of bias diminished the friction. We also show that the value of γ_0 as an integral of the memory is not in ad-equation with the observed values of the friction in the dataset. This is why we decided to ignore it for the rest of the study. The results of this measure are presented in Figure S-5.

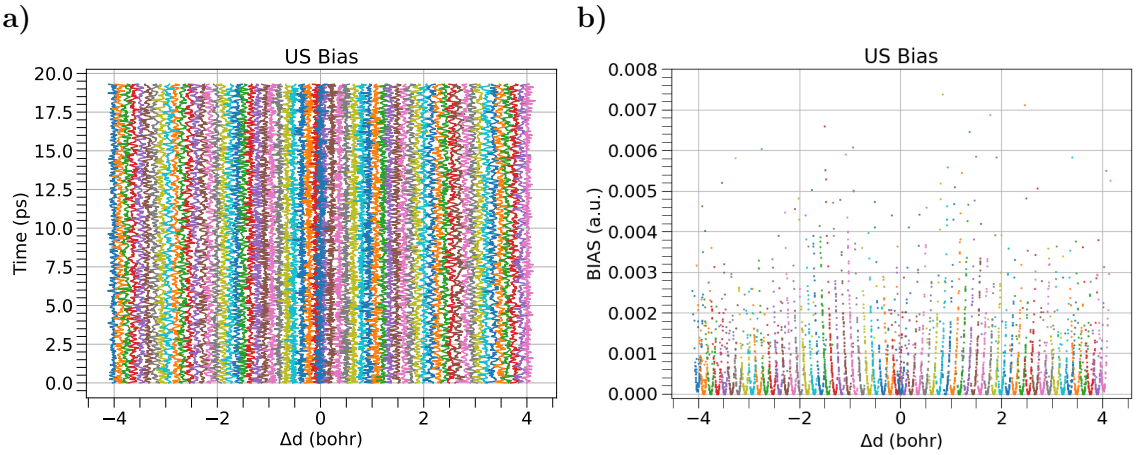


Figure S-4: Representation for time US data set for the MCCI system. **a)** representation of the time evolution of all the simulation windows. We can observe that the behavior of every windows remain stationary. **b)** The introduced bias with respect to Δd. In the reactants and products regions we can observe the effect of the additional external walls on cos(θ) and Δd.

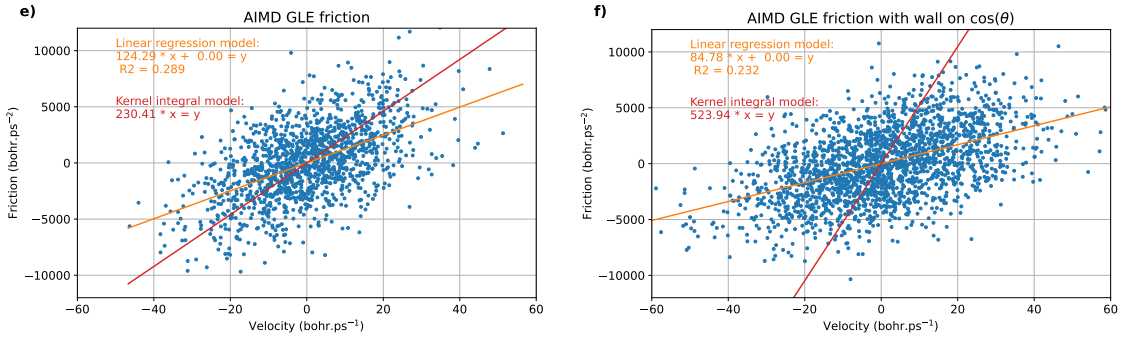


Figure S-5: Impact of the wall added on the cosine on the dynamics of the system. **e)** Estimation of γ_1 in the well ensemble with restrictive walls on Δd only. **f)** Same as panel e), with the additional wall on cos(θ).

S-9 Effect of the position on the memory

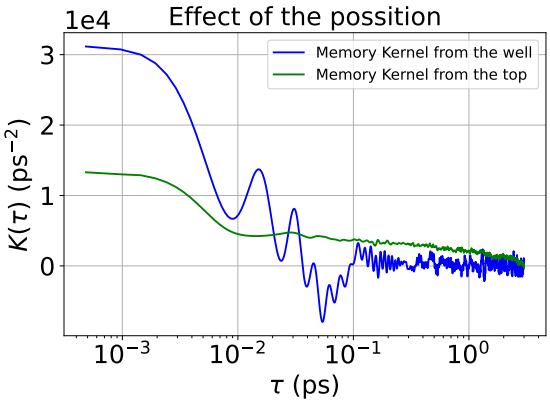


Figure S-6: Comparison of the memory kernel of the well with the one at the top of the barrier obtained with a bias to counter the barrier wall. The memory Kernel at the top present a long tail that could be an artifact generated by the bias.

In order to study the effect of the position on the memory kernel, we measured this at the top of the barrier by setting a quadratic bias that exactly compensates for the barrier, thereby sampling data at the top with the lowest possible amount of bias. The results of this study are presented in Figure S-6. The system's behavior at the top of the barrier is peculiar: the peak of the memory kernel at short times is lower and shorter than that for well data, but the remainder of the kernel exhibits a long tail that gradually approaches zero.

The presence of this long tail could be an artifact resulting from the bias, which introduces strong oscillatory behavior in the system that couples with other degrees of freedom. It is challenging to disentangle the effects of the bias from the effects of the position. It is possible that the long tail is entirely due to the bias, implying that the memory (and friction) is indeed lower at the top of the barrier than in the wells, which aligns with our intuition.

S-10 SI References

References

(1) Magrino, T.; Huet, L.; Saitta, A. M.; Pietrucci, F. Critical Assessment of Data-Driven versus Heuristic Reaction Coordinates in Solution Chemistry. *The Journal of Physical Chemistry A* **2022**, *126*, 8887–8900, Publisher: American Chemical Society.

(2) Daldrop, J. jandaldrop/bgle. 2024; `https://github.com/jandaldrop/bgle`, original-date: 2018-07-10T16:47:44Z.