Inventing general simulation methods to study the transformations of matter

Mémoire d'habilitation à diriger des recherches

Fabio Pietrucci

Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC)
Université Pierre et Marie Curie (UPMC), Paris

November 27, 2017
**Rapporteurs:**

**Razvan Caracas**  
Ecole Normale Supérieure de Lyon

**Christoph Dellago**  
Erwin Schrödinger Institute and University of Vienna

**Riccardo Ferrando**  
University of Genova

**Autres membres du jury:**

**Edouard Kierlik**  
Université Pierre et Marie Curie – Sorbonne Université, Paris

**Benjamin Rotenberg**  
Université Pierre et Marie Curie – Sorbonne Université, Paris

**Rodolphe Vuilleumier**  
Ecole Normale Supérieure et Université Pierre et Marie Curie – Sorbonne Université, Paris
Contents

1 Introducing the challenge 5
  1.1 Molecular dynamics and enhanced sampling . . . . . . . . . . . . . . . 6
  1.2 The role of reaction coordinates . . . . . . . . . . . . . . . . . . . . . 8

2 From the adjacency matrix to effective simulation tools 11
  2.1 Representing matter through the adjacency matrix . . . . . . . . . . . 11
  2.2 Social PeRmutation INvarianT (SPRINT) coordinates . . . . . . . . . 13
  2.3 Metadynamics based on SPRINT coordinates . . . . . . . . . . . . . . . 15

3 Structural transformations in the bulk 21
  3.1 Water poly(a)morphism: a simple metric resolves different topologies . 21
  3.2 Simulating transitions among ordered and disordered water forms . . . 24
  3.3 Not only bulk phases: application to nanostructures . . . . . . . . . . . 29

4 Focusing on molecular patterns: chemical reactions 31
  4.1 Path coordinates based on a coordination-pattern metric . . . . . . . . 32
  4.2 The network of reactions centered on formamide . . . . . . . . . . . . . 34
  4.3 Other applications . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

5 Outlook of future research 43
  5.1 Towards a deeper understanding and assessment of the new coordinates . 43
  5.2 Towards the efficient reconstruction of kinetic properties . . . . . . . 44

Bibliography 49
Introducing the challenge

“No matter what you look at, if you look at it closely enough, you are involved in the entire universe.”

– Michael Faraday

“Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.”

– Albert Einstein

This document describes an investigation started a dozen years ago, aimed at a deeper understanding of the ingenious and fascinating ways by which different atomic structures of matter transform in time. The research efforts, as often happens, followed a walk that is rather random, albeit with a clear drift toward the objective. Finding an effective balance between reachable, well-defined tasks and free creative detours is arguably one of the crucial dilemmas in pursuing scientific research – as well as in teaching it to younger collaborators. Along the way, we invented a handful new tools, based on the common idea of manipulating the matrix of inter-atomic connections as a proxy for the topology and geometry of the systems. The tools opened up to us new research directions, resulting in a palette of quite different applicative projects. Hence, the development efforts can be considered quite successful from an operative viewpoint. At the same time, the new tools are only partially accomplished, since no intensive work was done, so far, to explore their mathematical properties and to systematically optimize their performance. This will be one of our aims in the coming years. Another important perspective, as detailed in the last chapter, aims at a direct access to the kinetics of transformations, shifting the attention from time-less free-energy landscapes to the time evolution of rare events.
1. INTRODUCING THE CHALLENGE

1.1 Molecular dynamics and enhanced sampling

Condensed matter at the atomic scale transforms according to a multitude of mechanisms, sometimes counter-intuitive. A few examples – some will be discussed in the following chapters – are the nucleation of crystals from liquids, the formation of fullerenes from carbon fragments, the folding of proteins into a precise complex geometry, or the prebiotic reactions creating amino acids and nucleotides from much simpler molecules. The study of these transformation processes – it could be called the science of change – is a fascinating challenge, where computational methods play a major role due to the experimental difficulty in capturing evanescent transition states. The preceding examples of transformations are activated processes characterized by the following features: the system has different possible metastable structures – i.e., free energy minima separated by barriers – each one trapping the dynamics for a time that is long (easily seconds or more) compared to bond vibrations (of the order of femtoseconds). Eventually, a jump is performed toward another metastable state: note that the passage over the barrier is generally very rapid in itself.

In principle, molecular dynamics (MD) simulations are ideally suited to study the mechanisms, thermodynamics, and kinetics of activated processes at a given pressure and temperature. An extremely long trajectory of the system, generated from accurate forces, would ergodically sample fluctuations within metastable states as well as all possible transitions among them. However, \textit{ab initio} MD simulations (based on density functional theory) nowadays are typically limited to the sub-nanosecond time scale, whereas classical all-atom force fields are typically limited to the microsecond time scale. Hence, barriers due to breaking of one or a few interatomic bonds slow down the kinetics of activated processes to the point that standard computer simulations become of limited use. To cope with time-scale limitations, but also to get fundamental insight that is not automatically provided by a bare long trajectory, a number of different computational approaches were introduced in the last few decades, sometimes designed on purpose and sometimes borrowed from different disciplines.

An important point is the distinction between two problems: accelerated exploration and precise sampling. The former deals with the need to escape rapidly from
local minima. The latter consists in the extensive accumulation of samples in each relevant region of the configuration space, to estimate equilibrium probability distributions (and possibly kinetic properties). Some techniques are especially targeted to either of the issues, while others try addressing both at once. In general, the distinct nature of the two problems should be kept in mind, and, depending on specific system features, it can be more effective to tackle each task separately. The review Ref. discusses in some detail fundamental enhanced sampling concepts and selected applications. If we focus on algorithms based on dynamical trajectories at finite temperature, most existing methods exploit the following ideas:

1. couple the trajectory at the target temperature with trajectories at higher temperatures: the latter may diffuse more easily across barriers, since, according to Boltzmann distribution, the population of energy barriers relative to minima increases with temperature.

2. add to the natural forces in the system an artificial, external biasing force, designed in such a way as to enhance the population of high free energy regions (especially barriers) with respect to its negligible equilibrium value.

3. starting from a suitable configuration, breed successive generations of trajectories applying a sort of “natural selection” where only the offspring fulfilling some requirements is kept and propagated.

The studies described in this document exploited mainly the second class of algorithms: in this case, we consider a free energy landscape $F(s)$, defined as a function of a collective variable $s(x) \in \mathbb{R}^{3N}$ suitably tracking transformations, and enhanced sampling is achieved by adding a bias $V_B(s)$ the potential energy of the system. In a limiting situation, useful to illustrate the underlying idea, if $F(s) + V_B(s) \approx 0$ the landscape is flattened and barriers do not hamper any more the transitions (of course, in practice we do not know $F(s)$ in advance). More generally, we are free to build the bias with different forms and recipes, aimed at improving the exploration and/or sampling properties of the MD simulation. Some algorithms exploit a constant bias $V_B(s)$, others a time-dependent $V_B(s,t)$: analysing the results of the latter is typically more cumbersome. The first class includes umbrella sampling\textsuperscript{2}, hyperdynamics\textsuperscript{3}, accelerated MD\textsuperscript{4}, boxed MD\textsuperscript{5}, and more. The second class includes local elevation\textsuperscript{6}, conformational flooding\textsuperscript{7}, metadynamics\textsuperscript{8}, self-healing umbrella sampling\textsuperscript{9}, adaptive biasing force\textsuperscript{10,11}, targeted MD\textsuperscript{12}, steered MD\textsuperscript{13}, adiabatic bias MD\textsuperscript{14,15}, and more. It is important to stress that the efficiency of all these techniques depends critically upon the choice of the collective variable $s(x)$, as discussed in the next section. In the following chapters, umbrella sampling and metadynamics are often employed, hence their basics are described here.

In umbrella sampling, a well-defined function $V_B(s)$ is chosen before the start, aimed to focus the sampling in a predefined region of configuration space. A common example is a parabolic $V_B(d_{ij}) = \frac{1}{2}(d_{ij} - d_0)^2$ restraining an interatomic distance $d_{ij}$ within the neighborhood of a fixed value $d_0$. Repeating similar simulations for different $d_0$ values allows to sample thoroughly a wide range of distances, including, e.g., the breaking of a chemical bond. Note that reasonable starting configurations, describing a continuous pathway, need to be provided for each window: this is an important limitation, thus umbrella sampling is often used in tandem with other techniques for the discovery of transition pathways like metadynamics or steered MD. Eventually, the biased probability distributions are first corrected mathematically to remove the bias, and then combined together in a global estimate of the free energy profile using some variant of the weighted histogram analysis method\textsuperscript{16–21}. Umbrella sampling is
INTRODUCING THE CHALLENGE

a rather robust approach, widely used in different fields: it has the advantage that additional simulation can be easily added where needed, systematically improving the statistics, and that windows can be run simultaneously in parallel. On the other hand, applications are mostly limited to 1D or 2D free energy landscapes, the initial $V_B(s)$ often needs optimization to enforce overlap of $s$ distributions between windows, and atomic trajectories must be carefully inspected – a non-trivial task – to assess convergence.

In metadynamics, the form of the bias is not guessed in advance, rather $V_B(s,t)$ is constructed on-the-fly as a sum of Gaussian contributions (“hills”) added along the trajectory of the collective variable $s(t)$ at regular intervals. The Gaussians, small both in height and width compared to the local minima, tend to fill the latter bringing the system to overcome large barriers in a short simulation time. The basic idea is that, by definition, the system spends more time in low free energy regions, thus the local shape of the landscape can be approximately inferred from short-time probability distributions. The algorithm is usually most effective on low-dimensional spaces of collective variables (1D or 2D), so that the bias force acts in a precise direction. In some cases, however, high dimensional spaces are also employed, somehow more in the spirit of the local elevation approach. Metadynamics enjoyed a broad diffusion thanks to the simplicity of the algorithm and to the considerable amount of work that clarified theoretical foundations and convergence properties, with several effective variants introduced. Besides accelerating rare transitions, metadynamics also allows, in favorable cases, to reconstruct the equilibrium free energy landscape from $V_B(s,t)$, like a cast shaped from a mould. This result may appear puzzling given the out-of-equilibrium nature of the algorithm. The basic rationale is the following: if stationary conditions are observed, with a uniform probability in the $s$ space, the system is free to diffuse without being trapped by barriers, and this means that the free energy landscape is flattened by the bias potential.

1.2 The role of reaction coordinates

A key ingredient in the study of activated processes is the reaction coordinate. In a one-step process connecting two metastable states $A$ and $B$, it can be defined as the committor function $p_B(x)$ (or equivalently $p_A(x)$) that associates to each configuration $x$ in the full $3N$-dimensional space of the system the probability to reach $B$ before $A$ (or equivalently $A$ before $B$). Such probability can be computed launching from $x$ a large number of MD trajectories, each with different initial velocities, and counting how many reach $B$ before reaching $A$. The committor function varies smoothly between zero and one, assuming a value of 0.5 at the transition state. Therefore it is a suitable progress indicator for the transformation, including kinetic information about the fate of configurations, hence a good definition of reaction coordinate. The applicability is general, going from $T > 0$ down to $T = 0$ (where it reduces to the so-called intrinsic reaction coordinate), encompassing barriers of energetic and/or entropic origin, purely diffusive processes, etc. In practice, even if an explicit committor evaluation for all possible $x$ is utopian, committor analysis is a very useful a-posteriori tool: a limited number of configurations along a sampled reactive pathway can be tested, and identified as transition states if they lead to $A$ and $B$ with approximately equal probability. In practice, typically few tens of short trajectories (stopped once they relax to $A$ or $B$) are shot from each trial configuration, at a limited computer cost.

An important application of committor analysis is the validation of a putative reaction coordinate $s(x)$: after reconstructing the free energy profile $F(s) = -k_B T \log(P(s))$,
1.2. THE ROLE OF REACTION COORDINATES

Figure 1.2: Schematic illustration of committor probabilities (upper panel), as determined from short trajectories shoted from a configuration of the system, and examples of bidimensional free energy landscapes (lower panel). In the latter, the profile $F(s)$ for fixed $q = q^*$ is shown as inset, together with the corresponding distribution of committor probabilities: $q$ is a bad reaction coordinate in cases b and d. Figure taken from Ref. [32].

configurations drawn from the vicinity of the saddle point should display a committor distribution peaked around 0.5, otherwise one must conclude that $s(x)$ is a poor approximation of the ideal reaction coordinate (see Fig. 1.2). A famous example is NaCl dissociation in aqueous solution: the sole ion-pair distance, neglecting solvent degrees of freedom, gives an unsatisfactory committor distribution [31-39,41]. Committor analysis is increasingly employed not only in classical but also in ab initio MD simulations, for transition state- and reaction coordinate validation [42-45].

Several algorithms have been designed to identify optimal reaction coordinates for specific processes, usually requiring extensive simulation [34,35,46-55]. Knowledge, even approximate, of the reaction coordinate is extremely useful to enhance the sampling of the activated process, using some of the enhanced sampling techniques mentioned before. Despite the differences among biasing algorithms like umbrella sampling, metadynamics, adaptive biasing force, temperature accelerated MD, steered MD etc., they all are likely to be effective if the reaction coordinate is known. Conversely, a poor definition of reaction coordinate can seriously hamper all of those algorithms, leading to non-optimal transition mechanisms and preventing free energy landscapes to converge. Due to the complexity and variety of condensed matter, and in particular of its dynamic behavior, in practical simulations no free-energy calculation method can be guaranteed to converge in all circumstances [31, despite misleading messages.
sometimes appearing in the literature. Finally, one should appreciate that finding the optimal reaction coordinate is more than a technical issue: it often leads to a deeper understanding of the nature and driving forces of a transformation process, hence it is desirable even when enhanced sampling is not needed.

In real-world research projects, detailed information about transition pathways is not available beforehand, with the result that approximate definitions of reaction coordinates are often heuristically guessed drawing on literature, subjective experience, and intuition. Albeit non-rigorous, such approach can be successful when test simulations are used to critically revise the original guess in an iterative process. The large cost of such trial-and-error recipe (in terms of computer- and especially human time), together with the desire to shed light on fundamental features of transformation mechanisms, motivated our search for as-general-as-possible reaction coordinate definitions. The resulting developments, described in the following three chapters, were built around the unifying idea that matter transformations can be described as changes of a matrix containing information about the inter-atomic bond network. The final aim is to enable tackling in an effective, robust, and reproducible way a range of challenging problems: from chemical reactions to phase transitions, from the evolution of nanostructures to the folding and aggregation of proteins.
2

From the adjacency matrix to effective simulation tools

2.1 Representing matter through the adjacency matrix

We consider transformation processes occurring in condensed matter. For the latter, we adopt a broad definition including all structures formed by the aggregation of atoms: all sorts of molecules, nanostructures, liquids, and solids, without restricting to inert materials but including also biological systems. The nuclei and electrons in all these systems, of course, follow the same equations of motion, as well as (under appropriate conditions) the principles of thermodynamics and statistical mechanics. We will focus especially on atoms as basic units. From the standpoint of theory and simulation, an important and fruitful question can be formulated: is there a general way to characterize, in atomic detail, the wide range of possible transformation mechanisms of condensed matter? The question is not trivial since processes can be as different as phase transitions, chemical reactions, or protein conformational changes (see Fig. 1.1). One immediately imagines, building upon the basic Cartesian coordinates of atoms, many different mathematical representations, that however are tailored on one or the other system: examples are sets of distances, cell parameters (for periodic systems), coordination numbers, symmetry indicators (e.g., Steinhard order parameters), dihedral angles, root mean square deviations, etc. Is it possible to find a general representation, that furthermore is suitable not only for static structures but also to accurately track transformations? We attempted to make a step forward in answering the question, at the same time providing new useful simulation tools.

A possible starting point is to see an atomic structure as a network: atoms would be the nodes, and links would be defined based on proximity. Likewise, structural transitions become network transformations. The advantages of this view are its generality, encompassing many processes in condensed matter, and the availability of the powerful graph theory toolbox. In particular, a graph of $N$ nodes can be represented by a $N \times N$ adjacency matrix (also called contact matrix): $a_{ij} = 1$ if nodes $i$ and $j$ are connected, and zero otherwise. The “connection” between atoms is somehow an arbitrary concept: one can choose it to correspond to a chemical bond (ionic, covalent, hydrogen-bond, etc.), i.e., an energetic definition, or to be just an arbitrary property of close-enough atoms. Both choices have merits, and for its simplicity we will follow the second one (that often is compatible with the first). In practice, it is convenient to introduce a switching function $f(d_{ij})$ of the inter-atomic distance, decaying smoothly from one to zero in a chosen range of distances (an example is the Fermi-Dirac function, however the precise mathematical definition is
irrelevant). The range is often inferred by inspecting pair distribution functions, so that $f$ differentiates bonded atomic pairs from second-shell pairs. In passing from plain distances to a monotonic function of the distances, the information is equivalent, except when distances are larger than the cutoff, in which case we lose information. The resulting adjacency matrix is composed by real-valued elements $a_{ij} = f(d_{ij})$: examples for different systems are shown in Fig. 2.1. In this way, continuous changes of atomic structure are reflected in continuous changes of the matrix, hence from the matrix we can construct suitable topology-sensitive coordinates, able to track the transformations of matter in a rather general way. Besides an analysis tool, such coordinates can be also exploited within enhanced sampling methods to accelerate activated processes.

Adopting the preceding adjacency matrix as a basic representation of matter, we are limited by at least two hypotheses. The first is that we can safely neglect large distances between pairs of atoms, hoping that they do not play an important role or that they are effectively included in the shorter-range network structure (think to a system of points: having a large-enough set of distance-constraints, it should be possible to reconstruct the whole structure). The second hypothesis is that symmetry operations leaving the distances unchanged do not affect our ability to distinguish structures. One can easily build counter-examples for both, however they typically appear ad-hoc extreme cases. One relevant exception is given by enantiomers: molecules differing under a mirror operation are, sometimes, better considered separately (e.g., they can have different taste, or, more dramatically, a medical drug can be the mirror image of a poison). Even though such situations are not conveniently addressed, approaches based on the adjacency matrix encompass a very large range of problems.

Within its limited scope, this document will not dwell upon the broad field of biomolecular activated processes. However, it is important to mention that approaches similar to those presented here were successfully applied also in this field,
where we made a sizable number of contributions in the last decade.

2.2 Social PeRmutation INvariant (SPRINT) coordinates

This chapter provides a first example of topology-based coordinates: within the sandbox of nanoclusters isomerization, it illustrates the philosophy behind the invention and exploitation of such tools, and it motivates our further developments in the direction of phase transitions and chemical reactions. The reader will hopefully be indulgent if here the adjective “social” is attributed, somehow playfully, to coordinates sensitive to the network of neighbors surrounding a given atom.

The physical properties of nanoscale clusters strongly depend on their precise structure, which unfortunately is often out of reach of experimental techniques. Simulations can thus play an important role in exploring the thermodynamically relevant configurations. A related problem consists in predicting the distinct possible structures of complex organic molecules sharing a given formula unit. Moreover, understanding and predicting isomerization, association and dissociation is an important issue: as a general rule valid in several fields, structure prediction is typically “easier” and more exploited than the study of pathways and associated kinetic properties. Numerous computational strategies have been applied to determine low-energy geometries of atomic clusters, including simulated annealing, genetic algorithms, random search, basin hopping, and minima hopping methods. However, an extensive exploration of the potential energy surface with the above techniques is too expensive for non-empirical methods. A first screening is generally made within either classical-potential or tight-binding schemes, and only later a small number of candidate structures are selected and optimized at the DFT level or beyond. In this way, however, one risks to miss relevant low-energy structures. On the chemistry side, advanced first-principles searches for isomers of small molecules often exploit static calculations, e.g., using the “symmetry adapted stochastic search” or the “scaled hyper-sphere search method.” However, automatic approaches of this kind become rapidly impractical for low symmetry cases and beyond a handful atoms.

A more general and efficient methodology should allow the ab initio exploration of a large portion of the free-energy surface, and should lead to both thermodynamically relevant isomers and structural transformation pathways. To this aim we introduced so-called “Social PeRmutation INvariant” (SPRINT) coordinates. Being invariant under permutation of identical atoms, they avoid wasting time in the exploration of physically-equivalent minima in the free energy landscape. Following spectral graph theory, we consider the adjacency matrix $a_{ij}$ of the nanocluster, where the $ij$ are all pairs of atoms. The matrix is symmetric, non-negative, and also irreducible when it represents a connected graph, i.e., if any pair of vertices is connected through a path. In this case the Perron-Frobenius theorem holds: the largest modulus eigenvalue $\lambda_{max}$ is real, positive and non-degenerate, and the corresponding eigenvector $v_{max}$ has all non-zero components with equal sign. We adopt the positive sign convention. In particular, a few very interesting properties can be shown (to simplify, we consider here the classic binary adjacency matrix):

- $\lambda_{max}$ carries global information on the network: it grows with the number of bonds and lies between the average and the maximum coordination number;
- $v_{max}$ carries information about both the short- and long-range topology of the atomic network surrounding atom $i$: for any positive integer $M$

$$v_{i}^{max} = \frac{1}{(\lambda_{max})^M} \sum_j (a^M)_{ij} v_{j}^{max}$$

(2.1)
2. FROM THE ADJACENCY MATRIX TO EFFECTIVE SIMULATION TOOLS

Figure 2.2: Examples of simple graphs with each vertex labeled by the topological coordinate $S_i$ in Eq. 2.2. Clearly the coordinate of a vertex is sensitive both to the number of neighbors and to the connectivity of the neighbors themself.

where $(a^M)_{ij}$ is the number of walks of length $M$ connecting $i$ and $j$. Eq. 2.1 shows the “social character” of $v_{\text{max}}^{\text{max}}$.

These observations led us to combine the largest eigenvalue and corresponding eigenvector into the definition of topological SPRINT coordinates:

$$S_i = \sqrt{N}^{\lambda_{\text{max}}} v_{i, \text{max, sorted}}^{\text{max}}; \quad i = 1, 2, ..., N$$ (2.2)

where $N$ is the number of atoms, and the $i$th component must be taken after sorting the eigenvector from its smallest to its largest component. It is this sorting operation that makes the set $\{S_i\}$ invariant with respect to the $N!$ permutations of the labeling of $N$ identical atoms (and thus also with respect to point-group symmetries). Clearly, the constraint $S_1 \leq S_2 \leq ... \leq S_N$ strongly reduces the volume of the space to be explored. On the contrary, a permutation changes the order of rows and columns in the contact matrix $a_{ij}$, which therefore lacks this desirable invariance. An additional advantage is the substantial dimensional reduction from $N(N-1)/2$ elements of the contact matrix to $N$ elements only. The simple example in Fig. 2.2 illustrates how the SPRINT coordinates work, namely, how they single out the topologically inequivalent atoms and relate to both the coordination and the longer-range topology. Following the discussion in the previous section, in practical simulations we generalized $a_{ij}$ from a binary to a real-valued matrix by means of switching functions. In this case, due to the gradual decay of the switching function, the associated $S_i$ contain information not only on the cluster topology but also, to some extent, on the 3D geometry. SPRINT coordinates are not a rigorous solution to the problem of resolving isomorphic graphs from truly different ones: in other words, different graphs can share the same SPRINT coordinates. In practice, however, this degeneracy is probably rare in small, physically meaningful, nanoclusters, where so far these coordinates proved particularly useful. Instead, different crystal phases sharing a same number of first neighbors for all atoms (e.g., silica) cannot be efficiently resolved by this approach. A solution is presented in next chapter.

We implemented SPRINT coordinates in the plugin Plumed, allowing to perform free energy calculations with a number of classical and ab initio MD codes. It is interesting to note how the principal eigenvector of the adjacency matrix inspired remarkable applications in very different fields, including the compact description of protein structures, the centrality concept in social networks, and, last but not least, the tremendously successful PageRank algorithm of Google – based on the “25 billion dollars eigenvector.”
2.3 Metadynamics based on SPRINT coordinates

Based on the preceding considerations, we proposed to use the $S_i$ coordinates in computer simulations both for classification of nanostructure geometries and for efficient exploration of the latter, i.e., as collective variables for enhanced sampling simulations. Figure 2.3 shows how SPRINT coordinates lead to a quick exploration of Lennard-Jones (LJ) clusters, under the metadynamics bias, including global minima. Actually, since many variables are biased at once (one per atom), unconventionally for metadynamics applications, the algorithm is somehow in the spirit of local elevation\[6\]. The simulation rapidly visits the well-known lowest-energy isomers, i.e. icosahedra for $N = 13$ and $55$ and a face-centered truncated octahedron for $N = 38$ (Fig.2.3(b)). A crowd of isomers quickly emerge from the trajectory, in a sort of in silico experiment that does not take any input about the possible end-structures. In practice, starting from a given initial structure, a blind exploration is performed, where the system is forced to visit many different topologies. The spread of the $S_i$ values allows one easily to recognize highly symmetric from low-symmetry structures (see Fig.2.3(a)); more generally, during the simulation, collapse from a wide to a narrow range marks a disorder-order transition.

Figure 2.3: Metadynamics simulation of Lennard-Jones clusters at $T = 0.1$ (LJ units): a) $LJ_{13}$: time evolution of the SPRINT coordinates. Some of the explored isomers are shown. b) Lowest energy minima of $LJ_{13}$, $LJ_{38}$, $LJ_{55}$, visited within SPRINT-metadynamics trajectories. Colors vary from red to white to blue with decreasing values of $S_i$, helping to distinguish atoms surrounded with identical topologies.
Next, we consider the more realistic case of silicon clusters that were often taken as a test case for DFT-based algorithms designed to “search for the global minimum”. We exploited \textit{ab initio} Car-Parrinello\cite{77} MD at room temperature and in the local density approximation, to demonstrate that the Si’s allow, again, for a fast and efficient exploration of a multitude of low-energy isomers. Simulations started from fragments of simple cubic lattice, a far-from-optimal configuration. In the case of Si_{10}, the consensus lowest-energy structure (tetracapped trigonal prism) \cite{60,78} is quickly explored, among others. Increasing the cluster size, a rich variety of low-energy structures was observed. For example, in the case of Si_{16}, five independent simulations generated hundreds of different geometries over a cumulative time of 400 ps. Out of them, a set of eighty was selected and their atomic positions were optimized: sixteen were found within only 45 meV/atom as shown in Fig. 2.4. The lowest energy minimum agrees with previous studies\cite{79,80}. Already at this small size, an impressive diversity of structural motifs lie within a narrow energy range, from fused units (e.g., B, C, D, M) to capped (deformed) cores like the tri- or tetra-capped trigonal prism (A, E, F, G, I, K), from symmetric configurations to quasi-amorphous. All topologies are “inhomogeneous”, namely “highly social” (high-connected) atoms (red) coexist with “asocial” (low-connected) atoms (blue). By differentiating the atomic environment and identifying topologically equivalent positions, the SPRINT coordinates qualify as precise structural fingerprints.

In Ref.\cite{81}, we applied SPRINT-based \textit{ab initio} metadynamics to investigate the spontaneous transformation of small graphene nanoflakes into spheroidal cages, observed in transmission electron microscopy experiments\cite{82}. A set of simulations at different temperatures allowed to identify complex multi-step pathways, and discovered without any human bias ingenious transformation processes (Fig. 2.6). Among them, the zipping of a planar flake into a nanocone, the passage from a “bowl” to a

![Optimized geometries and relative energies for Si_{16} clusters, extracted from SPRINT-based \textit{ab initio} metadynamics simulations. Colors vary as in Fig. 2.3.](image)
2.3. METADYNAMICS BASED ON SPRINT COORDINATES

Figure 2.5: SPRINT coordinates distribution of the low-energy Si_{16} clusters.

cage with the help of carbon chains, and the formation of a pair of pentagons embedded into hexagons (the pyracylene unit, typical of fullerenes) by expelling a dimer from a four-membered ring. The latter transformation, in particular, is highly non-trivial, pointing to the exploration power of our technique, and is fully compatible with available experiments, being the inverse process of a well-known synthetic route to C_{62} from the C_{60} fullerene.

A different application of the same approach consists in exploring the possible isomerization reactions of organic molecules: e.g., in Ref. the possible topologies of a C_4H_5N molecule were explored, again with *ab initio* metadynamics at room temperature (Fig. 2.7). Remarkably, at times the molecule also splits in smaller fragments, that are able to recombine if a wall – i.e., a repulsive potential – prevents very large separations. Note that if the fragments remain within the range of the switching function, the adjacency matrix does not break down into separate blocks and the principal eigenvector components remain all positive. Note also that, since different elements are present, sorting of the principal eigenvector in Eq. 2.2 was performed only within sets of alike atoms. The simulation could explore a sizable list of linear and ring-like isomers, including pyrrole, and smaller molecules like acetylene and hydrogen cyanide. A similar approach, performed however in a more systematic way (analyzing an ensemble of trajectories), allowed Zheng and Pfaendtner in Ref. to
reconstruct a relatively complex reaction network for methanol oxidation, displaying analogies with the consensus network obtained from literature. Other SPRINT applications include the study of H$_2$SO$_4$ · HSO$_4^-$ dimer formation in the atmosphere, and the kinetics of the S$_N$2 reaction of CH$_3$Cl + Cl$^-$.

Simulations presented so far were performed in gas phase, i.e., for isolated systems. Another important class of systems is formed by molecules or clusters embedded into a solid matrix: in these cases, in addition to the internal degrees of freedom of the “guest”, its possible manifold interactions with the host have also to be included in the configuration space exploration. In Ref. 88, for instance, we applied DFT-based SPRINT-metadynamics simulations to identify the most probable configurations of B(OH)$_4^-$ (an important proxy of past ocean pH) inside calcite and aragonite crystals. Theoretical $^{11}$B NMR spectra of the predicted structures compare well with experimental results.

Besides enhanced sampling, a different, interesting use of topological coordinates concerns the automatic classification of molecular topologies. The basic idea is that different structural isomers are (usually) characterized by different SPRINT vectors, that can be directly compared to detect similarity thanks to their invariance under permutation of identical atoms. Examples are the study of complex combustion reactions of hydrocarbons, and the analysis of silicon nanostructures or of geometries explored in automated transition state search methods.

Figure 2.6:  a) From the 3D atomic structure of a graphene flake, a real-valued adjacency matrix is built, and its principal eigenvalue and eigenvector are combined together into SPRINT coordinates (for illustration purposes, in order to color atoms with the corresponding coordinates, no sorting is performed). b) Sequence of carbon nanostructures explored in an ab initio metadynamics trajectory, evolving into a spheroidal cage.
Figure 2.7: C$_4$H$_5$N: Two independent \textit{ab initio} metadynamics simulations employing the SPRINT coordinates, starting from (a) allyl cyanide and (b) pyrrole\textsuperscript{[23]}. 
A major challenge of materials science is the synthesis of new polymorphic structures with desired properties. In recent years, computational structure prediction methods\textsuperscript{93,94} have strongly contributed to the rapid increase of new predicted phases of materials, of high interest for potential applications (see, e.g., Ref.\textsuperscript{95}). However, at present, no general approach has been developed to guide experiments through the pathways connecting stable structures of condensed matter, so that synthesis remains a challenging endeavor, driven by trial-and-error and often dominated by kinetics rather than thermodynamics. In principle, atomistic simulations combined with enhanced sampling techniques provide a powerful tool for investigating transformation mechanisms and the corresponding energetics, but available approaches for phase transitions lack of transferability. In other words, a time-consuming system-dependent tailoring procedure is required for the study of each material, aimed at optimizing reaction coordinates (in simple cases, order parameters) and simulation setup. Moreover, it is not clear how to reach a polymorph very different from the available ones, thus possibly separated by several intermediate steps. Metastable phases are very often involved in phase transitions\textsuperscript{96} and sometimes their kinetic stability is very high. Thus, in order to recover the desired structure, one needs to find specific routes, e.g., acting on pressure or temperature, in a way that is not trivial to guess\textsuperscript{97}. A precise understanding of transition mechanisms and the corresponding kinetics is therefore the key to explain and control the behavior of matter. In this chapter a new technique is introduced trying to bridge the gap between structure prediction and synthetic routes prediction, initially demonstrated on water poly(a)morphism.

### 3.1 Water poly(a)morphism: a simple metric resolves different topologies

Despite the simplicity of its molecular unit, water is a challenging system, because of its uniquely rich polymorphism\textsuperscript{98}, the existence of different amorphous forms (polyamorphism), as well as predicted but yet unconfirmed features, including an elusive liquid-liquid transition\textsuperscript{99} and the possible formation of plastic phases\textsuperscript{100}. Here we show that using a novel metric, capturing changes in the topology of the interatomic network, we are able to differentiate the many possible structures. We further show that it becomes possible to systematically track transitions among liquid, amorphous and crystalline forms throughout the whole phase diagram of water, including the nucleation of crystals above and below the melting point.
Several experiments disclosed connections between stable and metastable water phases, while simulations highlighted the importance of metastable states in understanding the mechanism of phase transitions and related transformations. A classic example is the connection between the crystalline ice stable at ambient pressure (Ice I), and the low-density amorphous (LDA) and high-density amorphous (HDA) ices: by compressing Ice I up to 10 kbar at ≈ 80 K one obtains HDA instead of Ice VI, which may be transformed into LDA by decompression of HDA at 130 K or by heating recovered HDA at ambient pressure to beyond 130 K; finally Ice I is recovered by heating up LDA. Similar connections between crystalline and amorphous ices are found in the high-pressure region of the water phase diagram, where a very-high-density amorphous (VHDA) ice, plastic ices and crystalline structures with complex hydrogen-bond network (e.g. Ice VII) have been observed or predicted.

Molecular dynamics (MD) based on realistic interatomic potentials is in principle able to track such transitions. The kinetic barriers are, however, generally too large for an efficient exploration of the configuration space within the available timescale. To face this limitation, it has been so far necessary to introduce (i) simplistic models for intermolecular interactions and/or (ii) seeding techniques. Another approach consists in using enhanced sampling techniques, by focusing on low-dimensional collective variables (CV). Yet the CVs available to describe phase transitions are specifically designed for a given type of structural transformation, while no general CV scheme has been proven successful for a wide class of problems, in particular those involving amorphous systems. Recently, distance metrics developed for condensed matter have been proven to be successful in classifying structures in molecular or extended systems based on their atomic environment and/or interatomic network. In this respect, an important question is whether a given metric is able, besides classifying locally stable structures, to also track dynamical transitions in a continuous and accurate way. Here, a simple and general metric is discussed first, and in the following section it is employed to define a low-dimensional space of collective variables that is well-suited for enhanced sampling techniques.

In our algorithm, each configuration of the system is associated with a permutation invariant vector (PIV), built-up from inter-atomic Cartesian distances by re-organizing the elements of the (continuous) adjacency matrix, as in Fig. 3.1. The PIV is built starting from atom-type-specific ordered blocks, $v_{kk'}$, with elements

$$v_{kk'}^{\beta\beta'} = c_{kk'} S \left( \frac{\Omega_0}{\Omega} |r_{\beta k} - r_{\beta' k'}| \right).$$

Here $r_{\beta k}$ is the position vector of the $\beta$-th atom of type $k$ (oxygen or hydrogen), with $\beta > \beta'$, $k > k'$; $c_{kk'}$ are coefficients that define the PIV variant (they are all equal to one in the original formulation); $\Omega$ and $\Omega_0$ are the volume of the simulation box and a reference volume, respectively; $S$ is a switching function monotonically decreasing from one to zero as $|r_{\beta k} - r_{\beta' k'}|$ increases. In order to define the PIV, first the $v_{kk'}^{\beta\beta'}$ elements of each $v_{kk'}$ block are sorted in ascending ordered, then the different blocks are simply joined together resulting in a PIV of $N_{\text{at}}(N_{\text{at}} - 1)/2$ components, that we indicate with $V_\alpha$. The sorting operation within each block introduces invariance upon permutation of identical atoms. Note that the dimensionality of the vectors grows quickly with system size, so that sorting becomes computationally demanding and efficient tricks and parallelization are required (as implemented in the software tools piv-clustering and plumed). The volume scaling factor, absent in Refs., was introduced in Ref. to avoid violent fluctuations of the cell parameters during metadynamics simulations (see below). The choice of $c_{kk'}$ coefficients depends on...
3.1. WATER POLY(A)MORPHISM: A SIMPLE METRIC RESOLVES DIFFERENT TOPOLOGIES

Figure 3.1: Simple system of two water molecules illustrating the definition of PIV: the upper diagonal part of the $6 \times 6$ symmetric coordination matrix $C_{ij}$ is rearranged into a vector of 15 components $v_k$. The entries from each matrix block corresponding to O–O (in red), O–H (in yellow), and H–H pairs (in gray) are sorted in ascending order.

The decay range of the switching function is the main parameter entering the PIV definition, which is otherwise a very general one. Conceptually, the switching function $S$ is introduced to focus on a specified range of interatomic distances: short-to-medium range distances are typically included, i.e., a region displaying sizable differences between structures (for instance, at the level of pair distribution functions). A good practice is to evaluate the effect of different choices: in the case of water, we tested ranges going from about $2 - 5$ to $2 - 10$ Å. Typically, it is easy to find a distance range providing i) large Euclidean distances among PIVs of structures corresponding to different classes (e.g., liquid water vs several crystalline and amorphous forms), and ii) small distances (even including thermal fluctuations) among structures within the same class. Typically, such a range can be sizably increased or reduced without loosing the two properties above, even if the absolute values of distance can be strongly affected (e.g., compare Fig. 3.2 and 3.3).

A first interesting application of the metric consists in attempting to draw a 2D diagram, or map, that displays the Euclidean distances between PIVs of a set of structures. The dimensionality here is chosen just for ease of illustration. To build the map, a simple algorithm starts from random 2D positions of the representative points of each structure, and with a Monte Carlo procedure minimizes $\sum (d_{ij}^{\text{PIV}} - d_{ij}^{\text{map}})^2$, i.e., the difference between PIV distances and map distances. So far, different systems were analyzed in this way, including water (liquid, three amorphous forms, and several crystals), B$_2$O$_3$ (models of glass and polymorphs from Ref. [119], results unpublished), CO$_2$ (liquid and experimental polymorphs, results unpublished), SiO$_2$ (experimental polymorphs and glasses). In each of these cases, the simple 2D map could reproduce quite accurately (within errors of few percent) the distances among a dozen or more different crystalline and disordered structures, as represented by high-dimensional PIVs. This result is quite surprising (the reasons are unclear to us), and very helpful for the visual inspection of configuration spaces. Even more interesting is the fact that the 2D maps display, typically, an overall topology that is consistent...
Figure 3.2: a) Optimized two-dimensional map employing arbitrary axes accurately reproducing the PIV distances between water phases (Pearson’s correlation coefficient \( = 0.999 \)). PIV vectors contain here only oxygen atoms, with a switching function focused in the range 1–4.5 Å\(^2\). b) Experimental phase diagram of water\(^{113,117,118}\):

Note the analogies with the map. With the phase diagram: neighboring phases in the (P,T) space are also neighbor in the map of PIV distances, and amorphous phases are typically located at the interface between liquids and crystals to which they are directly connected in experiments. The example of water, in Fig. [3.3], illustrates well these considerations, that remain valid for the other systems mentioned above. This apparent correspondence between topological distances and kinetic proximity is suggestive of very useful applications: for instance, one could try organizing in such maps the results of powerful crystal structure prediction algorithms, gaining insight about pathways connecting an experimentally available phase with a desired new one. The next section demonstrates that the new metric offers also a relatively easy access to transformation mechanisms and free energy barriers, a further step towards the \textit{in silico} prediction of viable synthetic routes for new materials.

3.2 Simulating transitions among ordered and disordered water forms

Here we show that the PIV-based metric introduced in the last section allows to define a low-dimensional space of collective variables well-suited to perform enhanced sampling simulations. Such “topological” coordinates will be shown to accurately track the transformations among liquid, amorphous, and crystalline forms of water. The basic idea consists in combining the metric with so-called path collective variables\(^{121}\) given the structure of the system at a given time \( R(t) \) and a set of \( n \) reference structures providing a (discretized) putative pathway between an initial \( R_1 \) and a final \( R_n \) state, the variables are defined as

\[
s(t) = \frac{\sum_{k=1}^{n} ke^{-\lambda D(R(t), R_k)}}{\sum_{k'=1}^{n} e^{-\lambda D(R(t), R_{k'})}} \quad (3.2)
\]

\[
z(t) = -\frac{1}{\lambda} \log \left( \sum_{k=1}^{n} e^{-\lambda D(R(t), R_k)} \right) \quad (3.3)
\]

with \( s \) quantifying the progress of the transformation and \( z \) the distance from the putative path. Both are important: \( s \) allows to direct biasing forces to accomplish
3.2. SIMULATING TRANSITIONS AMONG ORDERED AND DISORDERED WATER FORMS

the transformation, whereas \( z \) discriminates between different pathways and allows tracking transitions to states different from the target. Note that the approach shares similarities with the string method \(^{122,123}\). The crucial ingredient is the metric \( D \) that here is taken to be simply the Euclidean distance between PIVs (see Eq. 3.1). \( \lambda \) is a parameter that can be conveniently set to be of the order of the inverse distance between neighboring reference configurations, assuming they are equally spaced. A much larger \( \lambda \) would produce very irregular and discontinuous pathways, while a much smaller one would hamper the resolution of different phases. In previous works, path collective variables were employed to study conformational changes of biomolecules and chemical reactions in gas phase. By introducing suitable new metrics \( D \), we demonstrated that these coordinates are extremely effective also for the simulation of phase transitions, transformations of nanostructures, and even chemical reactions with solvent participation (see next chapter). Furthermore, deviating from the original prescription of Ref. \(^{121}\), we posit that, within the latter classes of transformation processes, a putative \( n \)-membered path with \( n \gg 2 \) is in fact unnecessary, and it is possible to perform enhanced sampling simulations using only the initial and final states as references (\( n = 2 \)). This is a sizable simplification, since there is no more need for educated guesses about the transformation mechanism, which is discovered by the simulation itself, as demonstrated by applications.

In Figure 3.3-a we draw the pathways we followed on the phase diagram of a realistic model of water (TIP4P/2005\(^{124}\)), navigating within and across free-energy basins using standard MD and enhanced sampling techniques, respectively. Figure 3.3-b shows a two-dimensional map (see also previous section) of distances between the visited crystalline and amorphous structures: the metric employed to define the CVs is able to scatter the different phases in a way that recalls the topology of the phase diagram, representing kinetically connected phases as neighbors, and kinetically disconnected ones as far apart.

A few technical details: MD and enhanced sampling simulations were performed in the \( NPT \) ensemble. Due to the sizable computational cost of PIV construction, enhanced sampling simulations are an order of magnitude times slower than standard MD for PIVs including 800 water molecules, despite parallelization on 64 cores. Metadynamics\(^{8}\) simulations, of typical duration between 10 and 100 ns, allowed to overcome kinetic barriers and quickly explore many different transformations, discovering mechanisms as well as unexpected ice phases. To obtain free energy landscapes of high statistical precision we exploited umbrella sampling\(^{2}\) simulations, seeded from the structures explored with metadynamics. Each umbrella sampling window has a typical duration of 20 ns, with the last 5 ns (corresponding typically to more than 1000 times the autocorrelation decay time of the variables) employed to reconstruct the free energy landscape using the weighted histogram analysis method\(^{17}\). Statistical errors on free energies, estimated by bootstrapping or by cutting each trajectory in 5 segments and taking standard deviations, are smaller than \( 10^{-3}k_B T/N \). Gromacs 5.1.2\(^{126}\) and a modified version of the Plumed 2 plugin\(^{127}\) were employed.

As a starting point we analyze the crystallization of Ice I both from the liquid and the LDA phases at \( P = 1 \) bar and over a range of temperatures around the melting point (\( T_m \approx 250 \) K for the adopted interatomic potential\(^{124}\)). The initial configurations have been respectively obtained by cooling down an equilibrium liquid phase from \( T = 300 \) K and heating up a LDA structure from \( T = 100 \) K. Both crystallization transitions have been achieved multiple times in metadynamics simulations at \( T = 240 \) K and \( T = 260 \) K and they are all characterized by (i) a nucleation mechanism that we show in Figure 3.4 for the LDA–Ice I transformation at \( T = 240 \) K, (ii) the formation of a crystal nucleus of cubic symmetry (Ice Ic), and (iii) a final state with either a perfect cubic symmetry or made up of layers of cubic Ic and hexagonal
3. STRUCTURAL TRANSFORMATIONS IN THE BULK

Figure 3.3: a, The TIP4P/2005 water phase diagram is shown in grey, and phase transitions between (meta)stable phases (blue labels) simulated with metadynamics are indicated with red arrows. For selected transitions the detailed free energy landscape has been also computed with umbrella sampling (see text). Dashed green lines represent variations of the (P,T) conditions of the system within a phase, performed with unbiased molecular dynamics simulations. b, Two-dimensional map reproducing (within < 6% deviations) PIV distances between phases explored with MD and enhanced sampling. A switching function focused in the range 3–10 Å is employed, including both oxygen and hydrogen atoms. Axes are defined within an arbitrary rotation, here chosen so as to evidence the similarity with the phase diagram. See Ref. for details.

IH ice (see last snapshot in Figure 3.4-a). This last feature is in agreement with experimental findings, and our results show that the formation of stacking disordered Ice I may also proceed via the merging of Ic-nuclei as well as via (i) random growth of Ic and Ih layers and (ii) direct formation of Ic-Ih nuclei. In Figure 3.4-b we display the free energy profiles for the liquid–Ice I and LDA–Ice I transformations, respectively above and below the melting point. The calculated relative stabilities of the various phases agree with the phase diagram of the water model, and the order of magnitude of the free-energy profiles is consistent with free-energy differences calculated in previous work. We remark that crystallizing the liquid above the melting temperature, in the bulk, without any seeds and with a very realistic water model, shows that our approach allows to perform very challenging transformations even in unfavorable conditions, reaching metastable states (here Ice I) starting from the global minimum.

In order to carry on our continuous journey towards high pressure, we follow the experimental routes by cooling down Ice I to \( T = 100 \) K, and then compressing it at \( P = 10 \) kbar using standard MD: in the neighborhood of this point of the phase diagram we explore (with enhanced sampling simulations) the transformation to HDA as a function of temperature and pressure. We find that the free-energy barrier decreases when temperature and pressure increase, as we show in Figure 3.5. Furthermore we note that although the free energy of Ice I is higher than the one of HDA already at \( P = 10 \) kbar and \( T = 100 \) K (which is expected since Ice I is
3.2. SIMULATING TRANSITIONS AMONG ORDERED AND DISORDERED WATER FORMS

Figure 3.4: a, Sequence of snapshots ($\alpha, \beta, \gamma, \delta, \epsilon$) of the umbrella sampling simulation describing the progressive crystallization of LDA water into Ice I. Ice-like molecules, characterized by the oxygen having an averaged local tetrahedral bond order parameter higher than 0.7, are shown in blue (only for a single box replica). Snapshot $\gamma$ shows the Ice I nucleus at its critical size (saddle point in the free-energy profile). Snapshot $\epsilon$ shows stacking disordered ice I, with Ic cubic regions separated by Ih hexagonal layers (green lines).

b, Free-energy profile along the LDA-Ice I transformation pathway ($T = 240$ K, $P = 1$ bar) obtained via the weighted histogram analysis method applied to umbrella sampling trajectories. The $s$ and $z$ topological path collective variables represent the progress from the initial (equilibrated LDA) to the final (equilibrated Ice Ih) reference structure, and the distance from them, respectively. The obtained final crystalline state is a stacking disordered Ice I consistently with the relatively large value of the $z$ coordinate in the corresponding free energy basin (Ice Ih would be located at $s \simeq 1.9$ and $z \simeq 0$).

c, Comparison of the free-energy profiles, projected along the $s$ coordinate, for the crystallization transitions LDA – Ice I at $T = 240$ K and $P = 1$ bar and Liquid – Ice I at $T = 260$ K and $P = 1$ bar. In this latter simulation, the temperature is higher than the melting temperature ($T_m \simeq 250$ K for TIP4P/2005): Ice I correctly displays a free energy higher than the liquid phase. Both crystallizations were first achieved with metadynamics, followed by umbrella sampling simulations that yielded the profiles shown here. The free energy minima of each simulation are arbitrarily set to zero ($N = 800$ is the number of water molecules).

not the stable phase in this region of the phase diagram), the free-energy barrier is nonzero. This result is in accordance with a common interpretation of the Ice I–HDA transformation that can be seen as an extrapolation to low temperatures and high pressures of the Ice I–Liquid coexistence line. Even if this description is mostly qualitative, it certainly invigorates the idea that low-density crystalline ice is unstable with respect to denser disordered forms at high pressures, and that such instability occurs at higher pressures as temperature is decreased.
3. STRUCTURAL TRANSFORMATIONS IN THE BULK

Figure 3.5: a, Free-energy profiles projected along the path collective variable $s$ for the Ice I – HDA transformation at different thermodynamic conditions: $T = 100$ K and $P = 10$ kbar (black), $T = 120$ K and $P = 10$ kbar (red), $T = 100$ K and $P = 12$ kbar (blue). The free energy of HDA is arbitrarily set to zero. The inset shows the position of the three transformations in the phase diagram: at $T = 100$ K and $P = 10$ kbar Ice I is metastable (see Fig. 1-a) and separated from HDA by a barrier. Both variations $\Delta T = +20$ K and $\Delta P = +2$ kbar destabilize the crystal with respect to HDA making the free energy barrier disappear. b, Sequence of snapshots ($\alpha, \beta, \gamma$) of the system describing the amorphization process at $T = 100$ K and $P = 10$ kbar. They are taken from the umbrella sampling trajectories and their position along the $s$ coordinate is shown in panel a.

The HDA phase we obtain is then (i) decompressed at ambient pressure and transformed into LDA to close the loop of transitions at low pressure, (ii) compressed at $P = 12$ kbar and transformed into VHDA. The HDA–VHDA transformation connects the low-pressure and high-pressure regions that we explored (Fig. 3.3). The VHDA phase is compressed and heated until it reaches $P = 50$ kbar and $T = 300$ K, and at this point we address its crystallization to Ice VII. While simulating the VHDA–Ice VII transformation via metadynamics we observe that the system visits two additional metastable configurations. This result demonstrates that our method does not constrain the system to sample configurations along a simple path connecting the two reference structures, but rather allows it to follow complex mechanisms and discover new free energy basins. The metastable structures differ markedly from the other phases, as shown in Figure 3.6 where we compare representative snapshots and the oxygen-oxygen radial distribution functions. The first metastable phase is identified as the plastic Ice VII-P, which had already been proposed by Himoto et al.\textsuperscript{100} Oxygen atoms are arranged in a rather ordered crystalline network, whereas the hydrogen bond network changes dynamically: the correlation decay time of molecular dipoles is more than one order of magnitude shorter than that of Ice VII at the same thermodynamic conditions. The second metastable phase (labeled here “Ice Y”) is characterized by a tetragonal oxygen lattice and stacked layers of hydrogen-bond networks. A detailed investigation of this phase, including \textit{ab initio} simulations, will be the subject of a future work. The free energy landscape connecting the three crystalline phases Ice VII, Ice VII-P, and Ice Y (Fig. 3.6-c) illustrates the flexibility of the collective variables we designed. We remark that our approach is very general, hence of wide applicability to different materials, it allows tackling very challenging disorder-to-order transitions, and is highly complementary to structure prediction and materials discovery techniques. Ongoing applications include a more detailed
3.3. NOT ONLY BULK PHASES: APPLICATION TO NANOSTRUCTURES

Figure 3.6: a, Snapshots representing the oxygen atoms of the system in four different configurations: VHDA, Ice VII-P (plastic phase), Ice VII and Ice Y. All these structures are visited by the system while simulating the VHDA – Ice VII transformation at $T = 300$ K and $P = 50$ kbar via metadynamics. b, Oxygen–oxygen radial distribution functions as a function of the number of metadynamics simulation steps, for the stable configurations shown in panel a. The color code is consistent with panel a. c, Free energy landscape at $T = 300$ K and $P = 50$ kbar in the $\{s,z\}$ space obtained from umbrella sampling simulations. Equilibrated Ice VII and Ice Y are used to define the path collective variables: Ice VII-P appears as an unforeseen local minimum well-separated from the two reference crystals. The simulation box includes $N = 360$ water molecules.

investigation of water crystallization, a systematic study of its elusive liquid-liquid transition, as well as the reconstruction of free energy landscapes corresponding to polymorphism in CO$_2$ and B$_2$O$_3$.

3.3 Not only bulk phases: application to nanostructures

Despite the title of this chapter targeting bulk phases, we could recently employ with success path-PIV coordinates also in the study of complex transformation pathways of isolated nanostructures. Note that such systems have peculiar topological properties compared to infinite solids or liquids: atoms close to the nanoparticle surface are surrounded by a network that is, obviously, very different from the one in the core. Hence, the way of using the new coordinates could in principle need a modification: in facts, so far, it does not seem to be the case.

For gold clusters including $> 300$ atoms, it was possible, using empirical potentials, to simulate challenging transitions between icosahedral, fcc, fcc-twin and decahedral structures as observed in transmission electron microscopy experiments. In the case of even larger iron clusters, transitions between cube and Wulff-type forms were sim-
ulated, were iron atoms environments (bcc) do not change whereas the global shape of the cluster changes. In both problems, metadynamics and path-PIV coordinates including only the initial and final states as references work very easily, with results being robust with respect to the choice of the PIV switching function range. Two publications are under preparation. As of today, the approach appears very promising for a field where, similarly to solid materials, structure prediction is very well developed (especially with zero-temperature techniques) whereas transition pathways and kinetic properties are hardly accessible by simulation.

Finally, despite an early demonstration of the PIV-based metric for the clustering of chemical reaction trajectories in solution\cite{112}, to date the performance in combination with enhanced sampling has not been investigated in this context. In the future it will be interesting to compare this approach with the one described in the next chapter, where a metric specifically targeted to reactions delivered very satisfactory results on a range of problems, at the price of a less general formulation compared to PIV.
Focusing on molecular patterns: chemical reactions

A special class of transformation processes is characterized by very specific features, and is the object of a vast field of research and applications: chemical reactions. Most often, under this definition fall the rearrangements of chemical bonds experienced by one or more molecules, be it in isolation (gas phase) or in a crowded environment (a liquid solution or a solid surface, typically). Notwithstanding exceptions, typical features of chemical reactions are a sequence of discrete steps, and the localization of the changes both in time, within less than $10^{-12}$ s, and in space, involving a few atoms at once. Whereas the study of gas-phase reactions exploits well-developed theoretical and computational tools – the successful quantum chemistry techniques – the study of solvent and thermal effects is much less developed. This is due to the higher computational cost and to the more complex theoretical treatment of the latter problems. In this respect, MD is playing an increasing role, despite being typically limited to cheaper, less accurate quantum mechanical treatments (rarely beyond DFT) than in the gas phase.

The variety of atom types in the periodic table and of chemical bond types connecting atoms, create a vast palette of possible reactions. In particular, carbon skeletons combined with heteroatoms give rise to an infinite space of molecular architectures and related transformations, the object of organic chemistry. Remarkably, while sampling reactions from this immense potential reservoir, our planet managed to evolve simple molecules into complex biological molecules and, eventually, lifeforms. Prebiotic chemistry tries to understand the early steps of this evolution process: it is a growing, fascinating field, that inspired some of the studies described in this chapter. In this field, theory and simulation are still far from expressing their full potential, both in terms of explaining chemical reaction pathways (including the important role of different possible environments) and with respect to out-of-equilibrium thermodynamic foundations, until now quite elusive.

The aim of this chapter is to present the development and application of a set of topological coordinates aimed to track and simulate chemical reactions, once combined with ab initio MD and enhanced sampling tools. The starting point is again the adjacency matrix, recollecting the connectivity information about the system. At variance with the approach of the previous chapter, here the specific features of chemical reactions – particularly the localization of topological changes on a small number of atoms – are exploited to reduce the number of targeted degrees of freedom. The result is a simple approach that, so far, proved flexible and effective for a palette of different reactions. Some appealing features of the method are:
4. FOCUSING ON MOLECULAR PATTERNS: CHEMICAL REACTIONS

- the ability to deal on the same footing with gas-phase systems and with systems where the environment (solvent, solid surfaces, etc.) plays an active role;
- the inclusion of information only about reactants and (putative) products, without prior information – or prejudice – about the reaction pathway;
- the ability to automatically discover mechanisms, including unforeseen intermediates and products;
- the ability to obtain reversible reactive trajectories, a notorious challenge using traditional collective variables like distances and coordination numbers;
- the ease of use, compared with customary approaches that optimize collective variables by a cumbersome trial-and-error procedure, expensive in terms of computer- and human time.

4.1 Path coordinates based on a coordination-pattern metric

The widespread exploitation of MD simulations in the study of chemical reactions has been slowed down by the lack of general-purpose formulations of reaction coordinates. In particular, it is challenging to design coordinates that fully include the important role of the solvent degrees of freedom and that are general enough to be applied to a range of diverse reaction mechanisms. In Ref. we proposed a new way to tackle this problem. The method deals in the same formal way (namely, in the same space of coordinates) with gas phase and solutions, allowing a direct comparison of the two environments and an easier assessment of the effects of a given solvent on reaction pathways. The simulations described in the following typically exploit DFT-based Born-Oppenheimer MD simulations, with periodically repeated supercells including between a few (gas phase) and \( \sim 500 \) atoms (solutions).

Similarly to the previous chapter, the basic idea is to combine together two successful tools. The first tool is represented by path collective variables \( s \) and \( z \) (see Eqs. 3.2, 3.3). In the present context, umbrella sampling simulations of Branduardi and collaborators on methyl phosphate hydrolysis and on metadynamics simulations on \( CO_2 \) hydration and on transformations of carbon nanostructures clearly showed that path coordinates are a general and effective tool to reconstruct gas-phase \textit{ab initio} free-energy landscapes even for concerted reaction mechanisms, a case where simpler coordinates easily fail. In the latter works, however, the simple root-mean square deviation metric of Cartesian positions was employed to define the path coordinates: this choice requires to identify beforehand the role that each individual atom will play in the reaction mechanism. Moreover, multiple reference structures (a putative path) were employed to define \( s \) and \( z \).

Clearly, such demanding prerequisites are suitable only for relatively simple gas-phase reactions with a mechanism (including intermediates) that is not completely unknown. In general, it would be desirable to let the system “choose” which atom, among identical ones, will participate in a given reaction step (particularly, but not only, in solutions), and also which pathway is the most convenient one. We therefore introduced a different metric, based on the well-known fact that coordination numbers are a powerful collective variable for chemical reactions in gas phase and in solution. However, customarily only one or two coordination numbers are employed as landscape coordinates in free energy calculations, to comply with the practical need of a low-dimensional space. This is a severe limitation in general cases, where reactants, transition states and products display different values of several coordination numbers. We therefore decided to keep information about multiple (e.g., a dozen...
or even more) coordination numbers by defining the following distance $D$ between
the atomic configuration at time $t$ and the $k$th reference structure:

$$D(R(t), R_k) = \sum_{iS} (C_{iS}(t) - C_{iS}^k)^2$$

(4.1)

where $C_{iS}$ is the coordination number between individual atom $I$ of species (element)
$S'$ and all atoms $J$ of species $S$, defined as usual by a sum of smooth switching func-
tions decaying from one to zero for increasing distance. In practice, the topology of
the system is summarized into tables of coordination patterns, whose entries are the
sum over sets of entries of the all-to-all adjacency matrix. The metric takes into ac-

Figure 4.1: Construction principle of path collective variables based on coordination
patterns. The connectivity patterns of reactants and products are represented by
tables having individual, non-hydrogen atoms on rows and atomic species (the set of
all atoms of a given element) on columns. Arrows indicate changes of coordination
numbers, however all other matrix elements are free to change as well thanks to the
flexibility of path collective variables, a useful feature for the discovery of transition
mechanisms and unforeseen intermediate states.

count variations in a complex pattern of many coordination numbers during a reaction
mechanism (see Fig. 4.1), whereas its use within path coordinates reduces to only two
the dimensionality of the free energy landscape to be reconstructed. As demonstrated
by the following examples, this combination proved very effective for the simulation
of different type of reactions, even in the case where only two reference patterns (the
reactant and the putative product) are employed to define the coordinates. In this
case, remarkably, the system is typically able to discover a suitable reaction pathway,
including the possible appearance of locally-stable intermediate states.

The same definition of the metric $D$ takes into account gas-phase situations as
well as crowded environments like solutions and solids (or their combination, at a
liquid/solid interface). As an example, for reactions involving the participation of
one solute molecule in water, $I$ in Eq. 4.1 runs over several individual atoms of the
molecule (often excluding hydrogens, which can be exchanged rather easily), whereas solute and solvent atoms are contained in the groups $S$. Another convenient feature of the present formulation is that the detailed geometry of the reference structure needs not be explicitly given: instead, the matrix of $C_{IS}^k$ entries can be directly constructed starting simply from structural formulas (Fig. 4.1). In practice, employing the average $C_{kIS}^k$ values observed in a short MD simulation of each end-state can improve the resolution of intermediate and transition states in the free energy landscape. The reason is that smoothly-defined coordination numbers deviate from perfect integers.

We implemented the new coordinates in a modified version of the plugin Plumed 1.376 freely available on demand, compatible with DFT codes such as CPMD, CP2K and Quantum Espresso. Newer versions Plumed 2.x127 can be employed without modification, by exploiting suitable input structures. Plumed provides several enhanced-sampling and free-energy calculation techniques, and it can be easily interfaced with other codes, by adding a few calls from the MD engine.

4.2 The network of reactions centered on formamide

As a first application of the new coordinates, we reconstructed with ab initio meta-dynamics and umbrella sampling a set of chemical reactions involving formamide, HCONH$_2$.136,137 Increasing experimental and theoretical evidence points to formamide, ubiquitous in the solar system, as a possible hub in the complex network of prebiotic reactions leading from simple precursors like H$_2$, H$_2$O, N$_2$, NH$_3$, CO, CO$_2$ to key biological molecules like proteins, nucleic acids and sugars.137,138 Experimental setups, mimicking various origin-of-life scenarios, include laser sparks,136,139 UV light,140 proton irradiation,141 or shock waves akin to comet or meteorite impacts.136,142 Formamide can be formed from, and dissociated into, different fundamental building blocks, including H$_2$, H$_2$O, NH$_3$, CO, HCN, HNCO, HCOOH. Previous static calculations suggest that gas-phase barriers for these different processes lie in a relatively narrow range, providing synthetic flexibility.143,144 Given the possibly important role of water in prebiotic scenarios like surface- or submarine hydrothermal vents, parent bodies of meteorites, terrestrial impact sites of meteorites, etc., we also addressed formamide formation and decomposition in aqueous solution. Water can participate in formamide chemistry under different roles, including as a stabilizer through hydrogen bonds, as an efficient acid-base bifunctional catalyst, and as a co-reactant.145

In a prebiotic perspective, it is necessary a comparative understanding of reaction networks in different environments (gas or condensed phase, with different solvents and also interfaces with minerals) and at different conditions (temperature, pressure, irradiation, shock waves, etc.), eventually embracing also non-equilibrium scenarios, for their role in the emergence of life. For example, inspired by the classic Miller experiments, recent ab initio MD simulations demonstrated that an intense electric field can induce the barrierless formation of formamide and formic acid from a CH$_4$, CO, NH$_3$, H$_2$O, N$_2$ mixture in the condensed phase. Under the same conditions, formamide in turn gives birth to more complex organic molecules, including the simplest amino acid, glycine.146

In Ref.45 we simulated formamide decomposition and formation with ab initio metadynamics simulations (at DFT-PBE level with Grimme’s corrections).147 Both gas phase and aqueous solutions were investigated. Each reaction was simulated a few times modifying the hills size and deposition frequency, employing both standard and well-tempered metadynamics. Trajectories, of length between 50 and 200 ps, were examined and compared among them to assess the convergence of the free-energy landscape and estimate the statistical error bar (about ±2 kcal/mol). Overall, the
protocol adopted clearly showed that the new coordinates allow to discover realistic (by comparison with the literature) reaction pathways. Metadynamics proved very effective for the quick exploration of pathways, however the convergence of precise barrier heights proved more cumbersome to obtain, often demanding multiple simulations, each one encompassing multiple recrossings of the barrier. For this reason, in subsequent projects, we adopted a more efficient three-step protocol: i) a quick exploration of reaction pathways with metadynamics, followed by ii) committor analysis to ascertain transition states and unbiased pathways, and finally by iii) a reliable and precise reconstruction of free-energy profiles with umbrella sampling and weighted histogram analysis. This is an important practical point, since, arguably, the lack of general-purpose collective variables and of robust protocols yielding reproducible results are the key reasons, more than computer time, that so-far hampered systematic MD simulations of series of chemical reactions.

In a first set of simulations, at 300 K, we investigated formamide decarbonylation together with the inverse reaction:

\[
\text{HCONH}_2 \rightleftharpoons \text{CO} + \text{NH}_3
\]  

(4.2)

The average coordination patterns of equilibrated reactants and products provided the sole input for the construction of the \( s, z \) coordinates space, as described in the previous section. Within a single gas-phase simulation of 170 ps the trajectory explores 7 forward and 7 backward reactions. In the resulting free energy landscape (Fig. 4.2), the two basins corresponding to reactants and products differ by less than 1 kcal/mol in stability. Reactants and products are separated by a free energy barrier \( \Delta F^* = 79 \text{ kcal/mol} \), very similar to the barrier of 80.5 kcal/mol found in zero-temperature calculations with a more accurate quantum mechanical treatment (CCSD(T)/CBS+ZPE level) in Ref. 143. The transition state is also consistent with the latter study.

Figure 4.2: Free energy landscape for interconversion of formamide (A) and CO+NH\(_3\) (B) in the gas phase. Representative atomic configurations of free energy minima and transition states are shown as insets.

Compared to gas phase, the landscape of the solution results significantly modified in both quantitative and qualitative features (Fig. 4.3). In particular, besides the
focuses on molecular patterns: chemical reactions

Figure 4.3: Free energy landscape for reactions in aqueous solution between formamide (A), CO+NH$_3$ (B), and HCOOH+NH$_3$ (C). Representative atomic configurations are shown as insets.

reference states used to build the collective variables (formamide and CO+NH$_3$), the simulation explores a third basin featuring formic acid plus ammonia. This finding is remarkable since our space of collective variables was not explicitly constructed to take into account formic acid, that coherently appears at a high value of the $z$ coordinate with respect to the reference species (we recall that $z$ describes deviation from the references). Formamide, CO+NH$_3$ and HCOOH+NH$_3$ show a similar stability within our statistical uncertainty ($\pm 2$ kcal/mol). Inspection of the landscape results in a barrier of 35 kcal/mol for the reaction HCONH$_2$ $\rightarrow$ HCOOH+NH$_3$ and of 40 kcal/mol for the reaction HCONH$_2$ $\rightarrow$ CO+NH$_3$. The two decomposition channels, decarbonylation and hydrolysis, have thus quite similar barriers, confirming the effectiveness of topological path collective variables in discovering competitive pathways and even relevant chemical species. Additionally, ionic forms NH$_4^+$ and HCOO$^-$ are also explored, as expected from the basicity of ammonia and acidity of formic acid. An advantage of our molecular dynamics approach is to have built-in anharmonic temperature effects, so that we can directly break down $\Delta F$ into energetic and entropic contributions, as detailed in Ref.\(^4\). We remark that the different solute molecules interact in very different ways with water: formamide and formic acid are highly hydrophilic engaging on average in 4.5 and 5 hydrogen bonds with water, respectively, within an equilibrium MD simulation. Instead, CO appears as an hydrophobic molecule, not engaging in hydrogen bonds and featuring a first solvation shell at the sizable distance of 3.5 Å, similar to the hydrophobic cage of water molecules observed around solvated CO.

More in detail, all observed reactions begin with water donating a proton to the amino group of formamide, overcoming a barrier $\Delta F^* \approx 35$ kcal/mol. The resulting cation, that was identified as a relevant intermediate for this same reaction in Ref.\(^1\), is locally stable only within a time scale of less than 1 ps, as we verified with hundreds of short unbiased trajectories started from 10 different configurations featuring this
species. Experimentally formamide is a weak base, with the protonated species having a standard free energy almost equivalent (only 0.1 kcal/mol higher) to the neutral species, however the carbonyl is expected to be the most probable protonation site, in agreement with the transient nature of the protonated nitrogen in our simulations. Due to the transient nature and very short lifetime of the HCONH$_3^+$ intermediate, the mechanism could also be approximately considered as one-step. From HCONH$_3^+$ the system can evolve either towards carbon monoxide and ammonia or towards formic acid and ammonia. In the first channel, the dissociation of the C–N bond is accompanied by the donation of the carbon-bound proton to water. The mechanism appears fully reversible in our simulations, with backward transitions (formamide formation) featuring a barrier of 40 kcal/mol. The barrier can be flattened by a strong enough electric field as demonstrated in Ref. We employed committor analysis to validate the mechanism and assign transition state structures, shown in Fig. 4.4. Note that in the latter figure the transition states are not exactly on saddle points: presumably this is due to the aforementioned statistical uncertainty of the landscape, that we generally improved in subsequent works using umbrella sampling. In addition, in Ref. we demonstrated that introducing more than two reference structures for the construction of path coordinates allows, as expected, to improve systematically the separation among reactants, transition states, and intermediates on the landscape. It is interesting that both steps of the reaction, nitrogen protonation and carbon deprotonation, are enabled by the solvent, whereas in gas-phase the very same proton was transferred from the carbon to the nitrogen overcoming a higher barrier. We also remark the striking difference between committor trajectories in solution (Fig. 4.4), wandering in the low-curvature landscape regions for more than $\sim 100$ fs, compared to those in gas phase, resembling steepest descent relaxation pathways due to the lack of solvent collisions.

Figure 4.4: Examples of unbiased trajectories observed within committor analysis, and representative structures of transition states, for reactions among formamide, CO+NH$_3$ and HCOOH+NH$_3$.

The reaction channel corresponding to formamide hydrolysis

$$
\text{HCONH}_2 + \text{H}_2\text{O} \rightleftharpoons \text{HCOOH} + \text{NH}_3
$$

(4.3)
emerged (even if not explicitly sought for) as a decomposition pathway alternative to decarbonylation, showing the flexibility and predictive power of the new coordinates. Experimentally, at pH-neutral conditions the hydrolysis has a pseudo-first-order rate of $3.6 \times 10^{-9}$ s$^{-1}$ at 56°C\textsuperscript{151}, however the mechanism and free energy profile of the process have not been compellingly assigned. Based on static calculations including five water molecules Ref.\textsuperscript{152} claimed that hydrolysis would proceed preferentially through a water-assisted two-step mechanism, with the formation of an amino-gem-diol intermediate CH(OH)$_2$NH$_2$ (by adding water to the CO bond, the rate limiting step) followed by proton transfer from one OH group to the nitrogen and simultaneous dissociation of the CN bond. Previous Car-Parrinello steered-MD simulations of the aqueous solution\textsuperscript{153}, based on the C–O distance as reaction coordinate, suggested instead a concerted one-step mechanism with direct addition of water to the carbon in a tetrahedral transition state. Theoretical and experimental results were also interpreted in terms of a zwitterion mechanism. We also recall that the Miller-like \textit{ab initio} simulations in Ref.\textsuperscript{146} identified formic acid as an important prebiotic intermediate species, together with formamide, on a pathway towards the synthesis of glycine facilitated by an external electric field.

In our simulations the hydrolysis of formamide preferentially follows a two-step mechanism: first the nitrogen is protonated forming the transient HCONH$_2^+$ anion with a barrier of $\Delta F^* \approx 35$ kcal/mol (as discussed above). Inspection of the multiple reactive trajectories observed with metadynamics suggests that OH$^-$ addition to the carbon leads to the dissociation of the CN bond, thus completing the two-step reaction (albeit with a quite unstable intermediate). The backward reaction is also observed simply reverting the steps. Our barrier of 35 kcal/mol is similar to the one of Ref.\textsuperscript{152} and as shown therein in agreement with the available kinetic experimental data\textsuperscript{151}, however we remark how our two-step mechanism is different from the one proposed from the calculations in Ref.\textsuperscript{152} as it does not pass through an aminogem-diol intermediate (that we sporadically also observe off-pathway). The different result in Ref.\textsuperscript{152} could be related to the simplified model of solution (including five water molecules) employed by the authors. The one-step mechanism proposed from Car-Parrinello simulations in Ref.\textsuperscript{153} is also unlikely as it features a higher barrier (45 kcal/mol), possibly due to the simple collective variable employed by the authors, the C–O distance, including far less degrees of freedom than our topological path collective variables.

We then turned to investigate a different formamide decomposition channel, resulting in hydrogen (iso)cyanide and water:

$$\text{HCONH}_2 \rightleftharpoons \text{HCN} + \text{H}_2\text{O} \quad (4.4)$$

In gas phase, similar to before, we initially defined path collective variables based solely on the coordination patterns of formamide (reference state 1) and HCN+H$_2$O (reference state 2). However, whereas for decarbonylation we observed trajectories interconverting reactants and products without passing through intermediates, here the simulations revealed a more complex scenario. All reactions begin by the isomerization of formamide into OHCHNH through direct hydrogen transfer from nitrogen to oxygen. Next, the isomer evolves towards HNC$^+$H$_2$O, it reverts to a different OHCHNH isomer, and only at this point it reacts to form HCN$^+$H$_2$O. The main barriers are 45 kcal/mol for the initial internal proton transfer and 55 kcal/mol for the final dissociation giving HCN and water (to be compared with barriers of 42 and 59 kcal/mol, respectively, from zero temperature calculations at CCSD(T)+ZPE level\textsuperscript{143} and similarly at B3LYP level\textsuperscript{154}). In solution the inteconversion between formamide and HCN is still a multistep process (Fig. 4.5). The first step consists
4.2. THE NETWORK OF REACTIONS CENTERED ON FORMAMIDE

Figure 4.5: Free energy landscape for reactions in aqueous solution between formamide (A), OHCHNH (B), and HCN (C). Representative atomic configurations are shown as insets.

again in the formation of OHCHNH, however with the nitrogen-to-oxygen hydrogen transfer occurring through a proton chain in the solvent, sizably reducing the barrier to only 25 kcal/mol. Next, we observed a number of facile interconversions between the four different OHCHNH isomers, again thanks to proton exchanges with the solvent rather than dihedral rotations (that however are also present). HCN is then obtained overcoming a barrier of about 20 kcal/mol (or 45 kcal/mol taking formamide as a reference), the reverse barrier being about 44 kcal/mol. For comparison, barrier heights computed in gas-phase clusters including three/four water molecules at zero temperature and CCSD(T)+ZPE level are 21 kcal/mol from formamide to OHCHNH, similar to our value of 25 kcal/mol, and 35 kcal/mol from OHCHNH to HCN, sizably larger than our value of 20 kcal/mol; additionally, HCN is destabilized by 14 kcal/mol with respect to our results. According to committor analysis, the crucial reaction step is a proton transfer from the NH moiety to a water molecule, followed by the dissociation of the OH unit of the isomer and its protonation by the solvent to generate a water molecule. The process occurs through an hexagonal ring with the participation of three water molecules plus the newly generated one.

The free energy difference between formamide and HCN (plus water) is small, thanks to the favorable entropic term canceling out the unfavorable energetic one. The relative stability emerging from our simulations is in reasonable agreement with an experiments-based model of primitive oceans giving, at pH=7, a free energy difference of only 4 kcal/mol. This fact may be interpreted in the sense of both molecules having been largely available for possible prebiotic processes.

HCN appears as weakly hydrophilic, engaging on average in 0.5 hydrogen bonds with water and featuring a first solvation shell of radius 2.3 Å. Occasionally, overcoming a small barrier, HCN is also deprotonated to cyanide ion, at higher values of the z-path coordinate. Interestingly, at variance with the gas phase, in solution HNC is not locally stable, occasionally forming but then immediately disappearing within a timescale of ~ 0.1 ps. This observation is consistent with HNC being mainly found in interstellar clouds and only in gas-phase laboratory experiments, and it could
be interpreted in terms of HNC having a much more acidic proton than HCN.

An interesting question is how temperature and pressure modify prebiotic reaction networks, as different scenarios are characterized by different boundary conditions. As a first step in this direction we repeated gas-phase and liquid-solution simulations at $T = 400 \text{ K}$, keeping identical collective variables and simulation setup. The corresponding reaction landscapes, reported in supporting information of Ref. 45, show that in gas phase the effect on free energy differences is limited to within 3 kcal/mol, thus compatible with our statistical error bar of $\pm 2$ kcal/mol, whereas in solution both the transition state between OHCHNH and HCN and the local minimum of HCN are destabilized by 10 kcal/mol in passing from 300 to 400 K. The destabilization of HCN compared to formamide is traced back to an increase in $\Delta E$ (from 18 to 25 kcal/mol) concomitant with a decrease in $\Delta S$ (from 57 to 38 cal/(mol K)). The latter variation is consistent with an increased freedom of the formamide molecule and decreased freedom of HCN: the average number of solute-solvent hydrogen bonds passes from 4.5 to 3.6 for formamide and from 0.5 to 0.9 for HCN.

In a subsequent project we performed similar simulations to help rationalize experiments performed at the terawatt Prague Asterix Laser System, mimicking the effects of asteroid shock wave impact plasma, in addition to electric discharge, on a simple reducing mixture of NH$_3$, CO, and H$_2$O (a model of primitive atmosphere). The targeted process was the transformation of an atmosphere exposed either to the collision with an extraterrestrial body during Early and Late Heavy Bombardment periods in our solar system, or to lightning discharges in heavy clouds of dust, vapors, and other aerosols, resulting from impact, volcanic activity, and evaporation. The main outcome was the observation, among products, of all RNA nucleobases: their lack in previous Miller-Urey setups was one of the main criticism against the historical prebiotic experiment. Notwithstanding the limited scope of our Born-Oppenheimer MD simulations, aimed at thermally activated processes without explicit consideration of excited states and radical reactions, we could support the role of formamide as a suspected intermediate of reactions leading from simple prebiotic mixtures to nucleobases. In these simulations, free energy profiles were obtained using the efficient protocol mentioned before: quick metadynamics exploration of the mechanisms, followed by committor analysis to identify transition states and reactive pathways, followed by umbrella sampling started from unbiased committor trajectories.

### 4.3 Other applications

In conclusion, the new coordinates allow to study chemical reactions in gas-phase and in liquid-phase on the same footing, systematically obtaining reversible trajectories (a non-trivial result) and unveiling interesting mechanistic information without prior information. This step forward with respect to the state-of-the-art allowed our group to undertake the simulation of reactions in a handful of different systems, not only of prebiotic interest.

In Ref. 158 a liquid methanol model (32 molecules) has been simulated with ab initio MD (DFT-PBE) at 300 K. Application of an external static electric field of 0.3 eV/Å was shown to produce ions CH$_3$OH$^+$, CH$_3$O$^-$. Increasing the field intensity to 0.55 eV/Å leads to the spontaneous, barrierless formation of formaldehyde, methane and water. This disproportionation reaction, never reported before, could be interesting in the fields of energy and chemical synthesis. Note that electric fields of such intensity can be obtained, e.g., in electrode tip experiments and in a thin layer close to mineral surfaces. Metadynamics, committor analysis and umbrella sampling were employed at lower field intensities, where a reaction barrier appears,
4.3. OTHER APPLICATIONS

...to characterize in details the effect of the field on mechanisms. In Ref. [163], similar simulations suggested that dimethyl ether, an interesting alternative fuel, can be directly produced from methanol and accumulated for field intensities $> 0.55 \text{ eV/Å}$. In the latter works we discovered new reaction routes, bringing support to the emerging idea to use oriented electric fields as powerful tools in chemistry [164,165].

In two further projects of prebiotic flavor we could, once again, discover new reaction pathways and reconstruct the free energy barriers, consistently with experiments, by employing the coordination pattern-based coordinates. In the first, we addressed the decomposition mechanisms (decarboxylation and deamination) of two amino acids, glycine and isovaline, with the aim of rationalizing the relationships among amino acids and other organic molecules, observed in meteorites. In the second, we predicted – and confirmed experimentally by mass spectrometry – a new synthetic pathway yielding both purine and pyrimidine RNA nucleotides from phosphoribosyl pyrophosphate, the current biological precursor as well as a realistic prebiotic one, in simple hydrothermal water [167]. Detailed accounts of the results of both projects are in the process of being published.

Based on the different reactions studied so far, the new approach appears of general applicability. As a result, it now becomes envisageable the systematic study – with a unified enhanced-sampling approach and given DFT level – of a whole set of reactions in solution, carefully-selected to cover the main textbook classes and for the availability of experimental barriers. This is possible also thanks to the decreasing cost of trajectories of $\sim 100 \text{ ps}$ including $\sim 500$ atoms. The results of such study would allow to better assess the accuracy of DFT-MD simulations at finite temperature and in condensed phase, a missing information given that benchmark databases of reactions cover gas-phase zero-temperature processes.
5

Outlook of future research

Starting from the basic idea of watching atomistic structures as a network of atoms, this document presented three different approaches: SPRINT coordinates in chapter 2, suitable for small nanoclusters and isolated systems; PIV vectors and related path coordinates in chapter 3, effective for transformations both of bulk materials and relatively large nanostructures; path coordinates based on coordination patterns in chapter 4, targeted to reactions in gas phase and solution. In all cases, the starting point is the adjacency matrix of the graph, defined in a continuous fashion as a function of atomic coordinates. However, to achieve the best results on different classes of systems, the three approaches follow different recipes to extract information from the matrix. Coming back to the fundamental question put forward in section 2.1 – how to describe in a general way the transformations of matter – we only partially answered to it. We identified a possible general framework to describe most transformations of matter, but then, for reasons of simulation efficiency, we specialized to different recipes for different purposes. So far, PIV vectors appear as the most broadly-applicable and information-rich entities, able to keep track of topological changes in very different systems. There is some indication that they could be applied also to simulate chemical reactions in solution, making them, if confirmed, a very general tool, albeit of rather large computational cost (beyond a few hundred atoms).

The development efforts of the last decade put us in the position, today, of pursuing several applications. The latter span the fields of materials science, nanoscience, and chemistry: several projects were mentioned along the presentation. Additional research in biophysics, concerning the conformational dynamics of protein complexes and transmembrane channels, goes beyond the scope of the present document and will not be discussed here. Besides applicative projects, two method-development projects will be prioritary for our group in the next years: they are described in the following.

5.1 Towards a deeper understanding and assessment of the new coordinates

It is worth noting that the approaches presented in the last three chapters were invented in a very heuristic way, without resorting to sizable mathematical knowledge, tools, or justifications. Admittedly, the main reason is my lack of mathematical skills. A firmer understanding of the functioning principles and limitations of the coordinate spaces is highly desirable, even if the main objective remains the solution of problems
5. OUTLOOK OF FUTURE RESEARCH

of applied science. Future work will focus on a more rigorous assessment of the “quality” of the topological coordinates.

To begin with, we will extract from available databases a large set of crystallo-
graphic structures of different materials, including ionic, covalent, metallic and molec-
ular classes. All polymorphs – as well as amorphous models, where available – will be analyzed in terms of reciprocal PIV distance, to assess the ability of the metric to resolve different topologies (the distance should be larger than the spread of thermal fluctuations). If the excellent behavior observed for water, SiO₂, CO₂, and B₂O₃ will be confirmed, then we will be able to safely say that PIV distances appear a general approach. Recently we did a first step in this direction analysing 50 experimental polymorphs belonging to 13 different materials with bonding going from metallic to molecular, obtaining very encouraging results. Note that distance histograms have been proposed as structural fingerprints in Ref. 111,116, respectively for molecular databases (cheminformatics) and crystal structures. Our own definition differs from those in the latter works by the transformation of distances with a switching func-
tion (to focus on the neighborhood) and by the type of distance definition (Euclidean, summing contribution from each distance pair instead of each histogram bin). The operations of applying a cutoffed switching function and of ordering matrix elements, intuitively, suggest a loss of information with respect to the original distance matrix. Hence, it is likely that counterexamples could be mathematically constructed, with different structures resulting in degenerate PIV. in Ref. 108, for instance, different 4-atom models sharing the same set of 6 distances were put forward to conclude that distance histograms are not a complete representation of 3D structures. The real question is, however, whether such counterexamples have any true relevance for the space of physically-possible structures of materials, and in particular free-energy min-
ima. Our hope – based on the results so far on various systems – is that the answer is no, however a more systematic investigation is required and, of course, different definitions of structural distances will also be considered if they prove more effective.

Along another line, we plan to assess how well the collective variables we developed (both PIV-based and coordination pattern-based path coordinates) approximate the ideal reaction coordinate for a set of systems. The definitions could be then further optimized. Powerful techniques were developed in the last years for testing and optimizing reaction coordinates, including Refs. 34,46–55. We will select challenging processes, including chemical reactions in solution, phase transitions and nucleation events, and we will compare also with transition path sampling methods. 32,170–172. These techniques generate a large ensemble of realistic, unbiased transition pathways, providing a high-quality reference for comparing the results of our own approach. Whenever possible, clearly, we will also attempt a comparison with experiments, even if, from this point of view, dissecting the mechanism of transformations at atomistic detail is generally not trivial at all.

5.2 Towards the efficient reconstruction of kinetic properties

Enhanced sampling techniques, briefly sketched in the first chapter and reviewed in Ref. 1, alleviate to a large extent the time-scale limitations of MD: a large and growing literature shows that they yield qualitative and quantitative insight into many different transformation processes. In this context, the reconstruction of free energy landscapes is nowadays a mature field, where a number of successful strategies have been highly optimized and theoretically understood. An important challenge for the future is the discovery of systematic approaches to tackle the kinetics of complex systems in an affordable and accurate way. This is a large field of research on its own,
5.2. TOWARDS THE EFFICIENT RECONSTRUCTION OF KINETIC PROPERTIES

until now quite fragmented into different communities focused on physics, chemistry or biology.

At a fundamental level, time is key to transformation phenomena. Ideally, a complete understanding of an activated process would encompass all of its kinetic aspects, however such a description is often considered too difficult to achieve and one is content with reconstructing the geometric pathways and their free energy profile, i.e., formally time-less objects. In the near future, thanks to improving algorithms and computers, the reconstruction of kinetic rate networks is bound to gain the center of the scene, for several good reasons. Among them: i) The behavior of a number of crystalline or amorphous materials at relevant experimental conditions (in addition to all living systems) is dominated by kinetics rather than thermodynamics. In other words, the observed structure and composition of the system corresponds to a metastable state that, given a long enough time, would eventually cross a barrier and evolve towards a more stable state. Moreover, ii) a thorough understanding of the network of kinetic rates connecting the available metastable states, as a function of temperature, pressure, etc., is the key to enable the rational design of viable synthesis routes for complex materials or chemicals with desired structure and properties.

As it is well known, the synthesis of, e.g., advanced nanostructured materials or complex pharmaceutical molecules is often a matter of a delicate sequence of precisely-timed out-of-equilibrium actions (like heating quickly, suddenly adding a reactant, isolating an intermediate, centrifuging, etc.) driving the system on a narrow kinetic route. Presumably, similarly-tortuous processes led to the spontaneous emergence of biological molecules on primitive earth, and they recall, amusingly enough, the preparation of complex kitchen dishes. It is worth noting that the words “art” and “science” are very often evoked together speaking of sophisticated synthetic pathways, suggesting a lack of rational design.

Our plan is to address the development and validation of new approaches able to deliver the accurate transition times associated with different transformation mechanisms. We will exploit Langevin equations obtained from carefully-designed sets of short unbiased MD simulations. The idea stems from the cross-fertilization of concepts coming from distinct fields (crystal nucleation, dynamics of reactions in solution, protein folding). Several world-leading theoretical groups are dedicating growing efforts to develop kinetic modeling approaches, mostly in the domain of Markov state models but also, in a more limited way so far, in the direction of Langevin models. At present, Langevin equations are widely recognized as a key conceptual framework to discuss activated processes, nevertheless they are rarely employed as a computational device to pass from expensive all atom MD simulations to kinetic predictions. We aim instead at making Langevin models a practical tool to extract transition times from classical or \textit{ab initio} MD simulations.

The most popular kinetic tool in the biophysical community is represented by Markov state models, typically for protein folding. The configuration space is partitioned into hundreds of clusters, and the transition probabilities between such microstates are contained in a matrix of kinetic rates. The resulting first-order master equation is analyzed either by diagonalization (the largest eigenvalues and corresponding eigenvectors contain information on slow transitions) or by generating long stochastic jump sequences over the set of clusters (Gillespie's stochastic simulation approach, a.k.a. kinetic monte carlo). These models are typically built either from extremely long unbiased trajectories, available only in simple cases (folding of small peptides), or, in more complex systems, from biased (enhanced sampling) simulations. In this latter case this can be achieved using a technique we introduced where transition rates among microstates are given by an explicit formula, containing free energy differences and diffusion coefficients, derived by discretizing the Smoluchowski
Fokker-Planck equation. However, all these models are limited to memory-less and acceleration-free diffusion dynamics (corresponding to overdamped Langevin equations), a good approximation for proteins but not for other systems. Furthermore, even within this special case of dynamics, the results obtained from Markov state models sizably depend on the spatial coarse graining adopted and on the lag time (the time discretization interval used to analyze trajectories).

Langevin stochastic differential equations appear a more flexible and powerful tool to study the dynamics of general, complex systems. After having selected one or a few collective variables tracking the transformations, continuous trajectories in coordinate and momentum space are generated so as to faithfully reproduce the (projected) time evolution of the real system. Importantly, memory effects in friction and noise (the two being connected by the fluctuation-dissipation theorem) can be naturally included, leading to so-called generalized Langevin equations. The latter are known to be appropriate, e.g., for reactions in solution. The other two common types of Langevin equations, namely standard ones and overdamped ones, are just limiting cases of the generalized Langevin equation; the latter can be derived, at least in simple cases, from the full phase-space dynamics of a system using projection operators. There is some consensus that Langevin equations are a proper theoretical tool to describe activated processes throughout physics, chemistry, and biology. Given their low computational cost, Langevin models are amenable to easier extraction of mean first passage times and other kinetic quantities, with explicit analytic formulas being also known in selected cases.

To date, however, Langevin equations remain mostly a conceptual tool, and have been explicitly modeled on activated processes only sparsely in the literature, starting from extremely long ergodic trajectories – thus only ad hoc for simple problems like the conformational dynamics of small peptides. In these attempts, the availability of equilibrium MD trajectories spanning the whole phase space allowed to extract the drift, friction and noise fields directly as ensemble averages. Alternatively, the problem was split into the estimate of the diffusion coefficient from trajectories and of the free energy landscape from enhanced sampling methods.

The novelty of our project is twofold: i) we aim at a general approach, not limited to Markovian dynamics but encompassing also generalized Langevin equations, and ii) we will reconstruct at once the friction (hence noise) part and the free energy profile, without the need of extremely long ergodic MD trajectories or of separate expensive umbrella sampling-like simulations. Instead, we will exploit sets of short unbiased MD trajectories, in the spirit of committor analysis, keeping the computational cost to a reasonable level and opening to broad applications not only with classical MD but also with ab initio MD. If fulfilled, our effort will result in a powerful tool to reconstruct at once the thermodynamics and kinetics of complex systems.

We will develop and demonstrate our methodology on solutions, as paradigmatic systems featuring a high structural and dynamical complexity. We will address water solutions containing i) different ionic systems undergoing association/dissociation, from few-body processes to crystal nucleation, and ii) different organic molecules undergoing chemical reactions. Molecular dynamics simulations will be based on classical force fields in the former case and on density functional theory in the latter. The effects of different T, P and (if applicable) concentrations will be explicitly assessed. The processes will be chosen in order to cover different mechanistic behaviors and types of kinetics.

As a by-product of the studies, we will also progressively build up an online database of dynamical transition mechanisms (MD trajectories). This rather new concept fills a gap with respect to available databases that are limited to static geometric structures of crystal phases, nanoclusters, organic molecules, proteins, etc.
5.2. TOWARDS THE EFFICIENT RECONSTRUCTION OF KINETIC PROPERTIES

Hopefuly, the database will foster external contributions, collaborations and further investigations, and it will provide a set of useful benchmarks. In general, enhanced sampling and kinetic methods would benefit from the introduction of agreed benchmark systems of higher complexity than those customarily used (e.g., alanine dipeptide or small Lennard-Jones clusters), better approximating the rich behavior observed in realistic systems of interest. A significant example is the set of extensive MD trajectories of proteins generated with the Anton machine\textsuperscript{135} that helped reviving the race to efficient sampling methods for protein folding. Even if, recently, several groups attempted to critically assess and compare the performance of different enhanced sampling algorithms\textsuperscript{185–191}, a non-trivial task, still there is ample room for improvements.
Bibliography

BIBLIOGRAPHY


