Environmental and Nuclear Quantum Effects on Double Proton Transfer in the Guanine-Cytosine Base Pair

Federica Angiolari,[†] Simon Huppert,[‡] Fabio Pietrucci,[¶] and Riccardo Spezia[*,†]

†*Sorbonne Université, Laboratoire de Chimie Théorique, UMR 7616 CNRS, 4 Place Jussieu, 75005 Paris (France).*

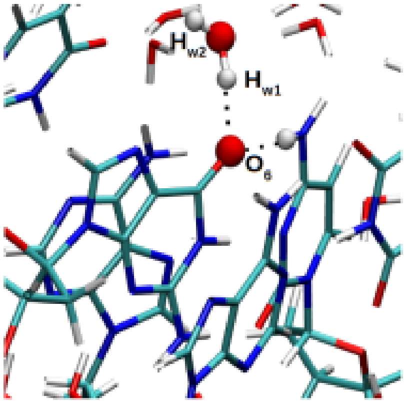
‡*Sorbonne Université, Institut de Nanosciences de Paris, UMR 7588 CNRS, 4 Place Jussieu, 75005 Paris (France)*

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

E-mail: riccardo.spezia@sorbonne-universite.fr

Abstract

In the present letter, we investigate the double proton transfer (DPT) tautomerization process in Guanine-Cytosine (GC) DNA base pairs. In particular, we study the influence of the biological environment on the mechanism, the kinetics and thermodynamics of such DPT. To this end, we present a molecular dynamics (MD) study in the tight-binding density functional theory framework, and compare the reactivity of the isolated GC dimer with that of the same dimer embedded in a small DNA structure. The impact of nuclear quantum effects (NQEs) is also evaluated using Path Integral based MD. Results show that in the isolated dimer, the DPT occurs via a concerted mechanism, while in the model biological environment, it turns into a step-wise process going through an intermediate structure. One of the water molecules in the vicinity of the proton transfer sites plays an important role as it changes H-bond pattern during the DPT reaction. The inclusion of NQEs has the effect of speeding up the tautomeric-to-canonical reaction, reflecting the destabilization of both the tautomeric and intermediate forms.



Base pairs mismatch is a phenomenon that occurs when the two DNA strands are not complementary. This means that the nucleotides which make up the base pairs do not match,[1] leading to an incorrect pairing in the double helix of the DNA. One of the possible mechanisms for this mismatch to occur, proposed by Lowdin,[2] is a double proton transfer (DPT) between the nucleotides causing an error during the replication of the DNA. Notably, DPT can cause a tautomerization of the base pair. Previous studies have shown that the Guanine-Cytosine (GC) pair is more prone to DPT than adenine-thymine (AT).[3,4] In the case of GC the most probable tautomer, as discussed in the litterature,[5] is that in which H_4 is transferred from N_4 of cytosine to O_6 of guanine and H_1 from N_1 of guanine to N_3 of cytosine (see numbering in Figure 1A with the corresponding canonical, GC, and tautomeric, G*C*, forms). Thus, if the G*C* tautomer is present during the replication, there is a possibility to form the non-standard G*T and AC* pairs, which in turn provides a possible explanation for the conversion of GC into AT,[6] as schematically shown in Figure 1B.

This DPT reaction, which is potentially responsible for tautomerism in base pairs, has raised significant interest in the theoretical community, while only a few experiments have been reported.[7,8] One detailed theoretical study of the reaction mechanism is reported by Ceron-Carrasco et al.[5], including the role of water micro-solvation using static DFT calculations. Notably, they found that the influence of the surrounding water molecules may change the mechanism from concerted to asynchronous, for the isolated dimer. Recently, Gheorghiu et al. studied the reaction pathways for tautomerism in GC and AT base pairs via QM/MM simulations, showing that GC can form the short-lived G*C* tautomer, while A*T* tautomerism was not observed.[4] Similarly, Li et al. employed a QM/MM approach to study tautomerism, focusing on wobble GT pairs.[9] Very recently a quantum mechanics/molecular mechanics (QM/MM) study by Soler-Polo et al.[10] using Umbrella Sampling[11] suggested that the water molecules and DNA environment destabilize the tautomeric form thus showing how nature has designed a robust base pair system. However, all these calculations do not consider the quantum nature of the proton, which is clearly an important aspect when studying proton transfer.[12–14]

The first study considering nuclear quantum effects (NQEs) on DPT in DNA base pairs is reported by Perez et al.[15] who combined Umbrella Sampling (US) with Path Integral molecular dynamics.[16] They found that the inclusion of NQEs clearly destabilizes the tautomeric form by flattening the free energy profile of the tautomeric state. However, they reduced the guanine-cytosine (GC) pair to a simpler model, taking into account only the atoms which take part in the mechanism. Slocombe et al.[17] have studied the tautomerism using DFT and Machine Learning Nudged Elastic Band methods with a tunneling correction to account for NQEs. They show that the G*C* tautomer has a lifetime long enough to survive during the cleavage process of DNA, while it is not the case for A*T*, which displays a very low reverse DPT barrier making the tautomeric form highly improbable. More recently, the DPT process in GC base pairs was also modelled using an open quantum system approach[18], suggesting that the tunnelling plays a central role even at biological temperature. However, all these studies that consider the quantum nature of the proton transfer, do not take into account the biological environment in the description of the proton transfer mechanism or only indirectly using a bath of harmonic oscillators to represent it.[18]

In the present study, we investigate the DPT reaction dynamics and thermodynamics including both environmental and nuclear quantum effects. We focus our study on the GC base pair since, as discussed previously, it is more prone to DPT than the AT base pair.[3,4,17] The reactivity is modelled using tight-binding density functional theory (DFTB)[19] which provides a good compromise between accuracy and computational accessibility, as was recently shown for different molecular systems.[20–22] Therefore in the following, electrons are always considered explicitly and quantum-mechanically, the label "classical" simulations refers only to the nuclear dynamics (as opposed to Path Integral simulations that include NQEs).

We consider two models of GC base pair: (i) isolated GC, (ii) GC embedded in a DNA model composed of three base pairs with GC in the middle, denoted hereafter 3BP-DNA. The embedded model was extracted from the 1D28 PDB structure[23] and corresponds to a TGA (Thymine-Guanine-Adenine) sequence. For this last model, the crystallographic water molecules from the X-Ray structure were also included. This corresponds to a micro-solvation model that is computationally allowed in conjunction with DFTB and path integrals and can be compared to some previous literature results[4,5]. Note that all the atoms are free to move in the simulations, including the crystallographic water molecules. The different systems studied are shown in Figure 1.

TGA is only one of the possible DNA sequences and in principle the nature of the bases above and below the GC pair could affect the reactivity. However, Cerón-Carrasco and Jacquemin[24] studied all the possible DNA-trimers and found that the proton transfer energies differ only by a very small amount. We have thus chosen to use a structure extracted from crystallographic data of an actual DNA sequence. From this sequence we extracted the aforementioned TGA trimer as well as a longer pentamer, TTGAG, used for validation of the mechanism (5BP-DNA).

Since, the DPT occurs between the canonical GC and the tautomeric G*C* forms shown in Figure 1A, four characteristic distances can be used to describe the reaction: $r_1 = |O_6 - H_4|$, $r_2 = |N_4 - H_4|$, $r_3 = |N_1 - H_1|$ and $r_4 = |N_3 - H_1|$. Two collective variables are then typically used for DPT reactions:[10,25,26] $d_1 = r_2 - r_1$ and $d_2 = r_3 - r_4$. Note that the G*C* form is typically considered in the literature as the most relevant tautomeric form from all the possible configurations in alternative to the canonical (GC) form.[5]

We first consider how the G*C* tautomer evolves dynamically once it is formed. To account for NQEs, we perform Ring Polymer Molecular Dynamics (RPMD) simulations starting from G*C*. The RPMD approximation is based on the Path Integral formalism[27,28] and it is known to correctly describe proton transfer at room temperature[25]. It was also recently used to study the stability of base pairs and the influence of NQEs on hydrogen bonds in DNA base pairs.[29]

Note that a combined theoretical and experimental work proposed that the tautomeric form is accessible photochemically:[7] once formed in the excited state it can eventually relax back to the canonical form. It is thus useful to understand the dynamics of the reaction pathway connecting the tautomeric form to the canonical form and the impact of NQEs and of the biological environment on this process.

Atomic interactions are modelled via DFTB using the MIO set of Slater-Kostner parameters[30] modified to better describe N–H bonds (MIO:NH parametrization).[31] The third-order expansion was used and dispersion was added at the D4 level.[32] This specific set of parameters was chosen after a thorough comparison between DFTB and high level electronic structure calculations reported in detail in the Supporting Information (section S1). Indeed, the level of theory employed can clearly modify the energy profile of the reaction and consequently its mechanism and rate constant.[4,9,17,33] We first compared the formation energy obtained with DFT to high-level calculations from the literature[5] and found that the CAM-B3LYP functional[34] with dispersion corrections[35] reproduces very well the interaction energy

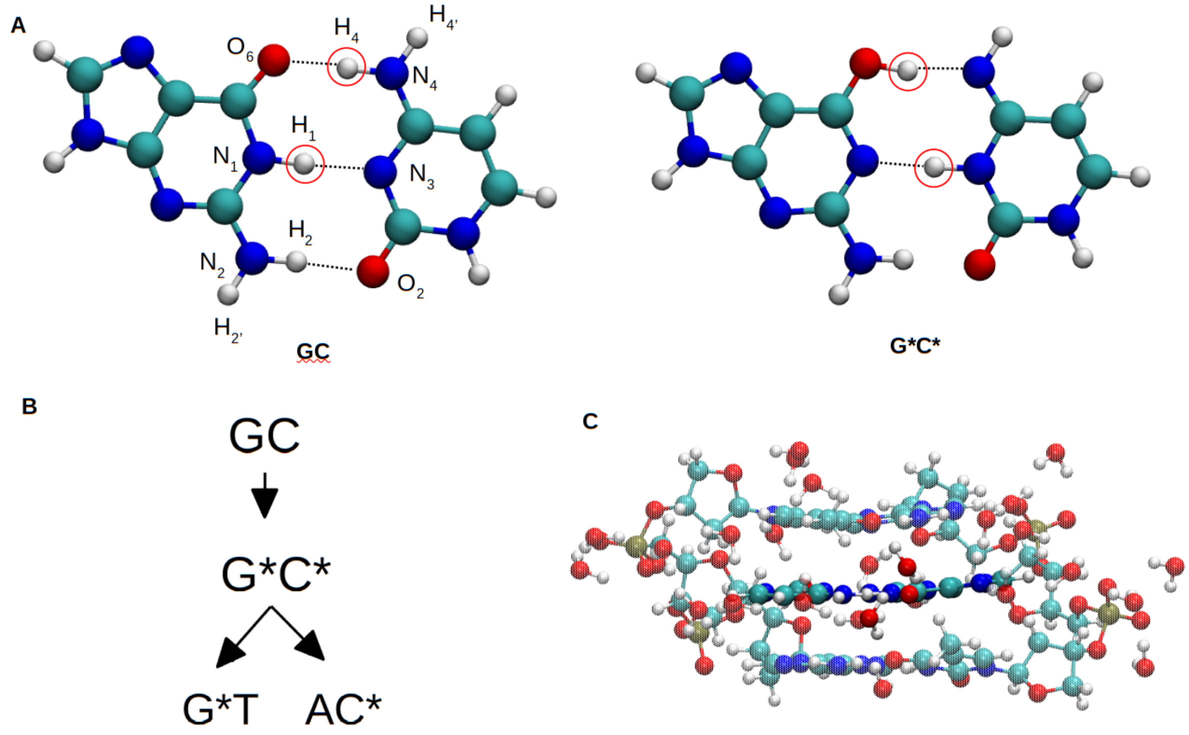


Figure 1: Panel A) Canonical (GC) and tautomeric (G*C*) forms of the Guanine-Cytosine base pair, with atom numbering. Transferring hydrogen atoms are shown in red circles. Panel B) Schematic representation of the tautomerism and its impact on the replication of DNA, with a GC base pair forming the G*C* tautomer and eventually leading to non-standard pairing G*T or AC* (A = Adenine and T = Thymine). Panel C) Structure of the 3BP-DNA model extracted from the 1D28 PDB structure[23] with the corresponding crystallographic water molecules.

and the equilibrium geometry (see Table S1 in the SI). Considering DFTB, we noticed that MIO:NH and OB2 parametrizations only slightly underestimate the dimer interaction energy (compared to other sets of parameters that yield larger errors), and provide good predictions for the geometry. To address the DPT reaction more specifically, we compared the energy profile of such reaction as obtained from the CAM-B3LYP DFT functional and different DFTB parametrizations, as shown in Figure S2 of the SI. Notably, MIO:NH provides a potential energy surface (PES) that agrees well with the DFT, while other parametrizations yield incorrect barrier shapes or much higher barriers. While the barrier (see Table S2 in the SI) is still slightly overestimated with the the MIO:NH parametrization, the shape of the PES is similar to that of CAM-B3LYP. Therefore even if the absolute value of the rate constants can be affected, the mechanism should likely not be impacted.

To study the DPT process, reaction dynamics simulations are initiated in the tautomeric form (G*C*) and propagated on the DFTB Born-Oppenheimer surface. Nuclear quantum effects are included within the RPMD framework using a Langevin thermostat with optimal damping on the fluctuation modes of the ring polymer, corresponding to the Thermostatted RPMD (T-RPMD) algorithm,[36] initially developed to improve vibrational spectrum calculations and later applied to reaction rates.[37] An additional thermostat on the centroid is added with a friction parameter of 10 ps $^{-1}$. Simulations are performed at 300 K and the number of beads (P) is set to 8. This value is lower than the one used in previous studies on analogous systems and may be not enough at room temperature,[15,25] but allows to capture the main impacts of NQEs. Further increase of P only causes a limited modification of the DPT rate constants while keeping the mechanisms unchanged, as reported in the Supplementary Information and briefly discussed below. Classical simulations (Langevin MD, LMD) are obtained setting $P = 1$. An ensemble of 100 trajectories is performed for each system, for both T-RPMD and LMD simulations. The simulation length is chosen in order to ensure that 100% of the G*C* structures react. Note that in this set of simulations, we do not force any reaction pathway: in this way, we can observe where the system naturally evolves, and it could therefore form one or more final structures through potentially different mechanisms.

We also computed the free energy surface for the DPT process via Umbrella Sampling (US)[38] using the d_1 and d_2 collective variables (CVs) previously defined. Notably, we have considered, for each CV, a total number of six windows between -0.75 and +0.75 Å, for a total of 36 points on a bi-dimensional grid. Each run was of 5 ps length with a time-step of 1 fs. We used the standard Weighted Histogram Analysis Method (WHAM)[39] to obtain the free energy surface, and uncertainties were estimated using block-average analysis. Further details are given in the Supporting Information, section S2. The US was performed for both classical and Path Integral MD (PIMD) simulations, with 8 beads: in this second case, we used the reaction coordinates defined from the centroid positions, as suggested by

Hinsen and Roux for a general proton transfer reaction[40], as well as other studies on reactive dynamics.[28,41]

The DFTB energies and gradients are calculated with the DFTB+ software (version 22.1),[32] and the T-RPMD and LMD simulations are implemented via our own in-house code recently detailed and tested on simple analytical potentials.[42] We have presently interfaced it with DFTB+ to study large molecular systems. US simulations are performed using the Plumed software[43,44] as a library imported in our T-RPMD and LMD code. More details on trajectory simulations are reported in the Supporting Information (section S2).

All trajectories initially in the G*C* tautomeric form spontaneously end up in the canonical form during our simulation time length. However, the reaction rates and mechanisms are dramatically affected by both NQEs and the DNA environment. As discussed previously, two collective variables are typically used to describe the DPT, d_1 and d_2, where d_1 corresponds to the external proton that interacts more strongly with the surrounding water molecules (when present). In Figure 2, we show the projection of the direct dynamics on the d_1-d_2 plane (the trajectories evolve in the full-dimensional phase space without any constraint), as obtained for the isolated base pair and for the 3BP-DNA system. In both cases, classical (LMD) and quantum (T-RPMD) results are shown. The GC canonical form corresponds to negative values (the minimum is around d_1 = -0.7 Å and d_2 = -0.8 Å), and the tautomeric G*C* form corresponds to positive values (with a minimum around $d_1 = 0.7$ Å and $d_2 = 0.7$ Å).

For the isolated system, the DPT occurs in a concerted way (i.e. along the diagonal in the d_1–d_2 plot of Figure 2) similarly to what was previously suggested.[26] As it can be noticed, there is a slight deviation from the exact diagonal that roughly corresponds to a slightly asynchronous mechanism, but it disappears when taking into account NQEs. Indeed, the inclusion of NQEs has the effect of significantly accelerating the DPT reaction and of making the mechanism fully synchronous. From direct dynamics simulations it is possible to evaluate unimolecular rate constants (k) and the corresponding life-times ($\tau = 1/k$) from an exponential fit of the population decay of the initial state (here the G*C* tautomer), as shown in Figure S3 of the SI. The rate constants and corresponding uncertainties, obtained via the bootstrap method as in our recent work,[42] are listed in Table 1. The G*C* rate constant of the isolated system estimated from the T-RPMD trajectories (2.4 ps $^{-1}$) is about 30 times greater than the LMD one (0.0821 ps $^{-1}$). This result is in agreement with Perez et al.[15] who shows that the barrier in the free energy pathway connecting G*C* with GC almost disappears when including NQEs, using Path Integral Umbrella Sampling. Note that, when increasing the number of beads (see Table S7 of the SI) the rate constant further increases from 2.4 to 4.0 ps $^{-1}$, but the mechanism does not change, showing that 8-beads results are not totally converged for the isolated system but already capture the correct trend. When we deuterate the system, by substituting the two transferring H atoms with D, the impact of NQEs is reduced, as expected. We obtain a ratio $k^H/k^D = 3.8 \pm 0.8$, while this value is 1.2 ± 0.1 in LMD simulations.

Table 1: Comparison between the rate constant (k) of the reverse reaction (from tautomeric to canonic) for the two systems: isolated GC dimer and the 3BP-DNA model as obtained from LMD and T-RPMD (using 8 beads) DFTB-based direct dynamics simulations. We report also the values obtained for deuterated systems. Ratios between LMD and T-RPMD as well as hydrogen (H) and deuterium (D) rate constants are also shown.

System	k $[ps^{-1}]$	k^{T-RPMD}/k^{LMD}	k^H/k^D
Isolated/LMD	0.0821 ± 0.009	–	–
Isolated/LMD/Deuterated	0.070 ± 0.008	–	1.2 ± 0.1
Isolated/T-RPMD	2.4 ± 0.2	29 ± 5	–
Isolated/T-RPMD/Deuterated	0.64 ± 0.07	9 ± 1	3.8 ± 0.8
3BP-DNA/LMD	1.7 ± 0.2	–	–
3BP-DNA/LMD/Deuterated	1.3 ± 0.1	–	1.3 ± 0.2
3BP-DNA/T-RPMD	18 ± 2	11 ± 2	–
3BP-DNA/T-RPMD/Deuterated	6.9 ± 0.9	5 ± 1	2.6 ± 0.5

When the model DNA environment is taken into account, the DPT mechanism clearly changes: we now observe a step-wise mechanism in which H_4 (the less acidic proton) moves first from O_6 to N_4, forming an intermediate structure and then, in a second step, the other proton (H_1) moves and finally forms the neutral canonical form (GC). This intermediate structure is described in more detail below.

The inclusion of the DNA model environment also has the effect of further destabilizing the G*C* tautomer, as shown by the corresponding rate constants reported in Table 1. This result is in agreement with recent QM/MM simulations showing that the tautomeric form is thermodynamically destabilized when a DNA-like environment is included.[10] The acceleration due to NQEs is slightly reduced compared to the isolated dimer, as k^{T-RPMD}/k^{LMD} is lowered down to 11. Isotopic substitution results also reflect this finding as k^H/k^D is now reduced to 2.6. Interestingly, when performing T-RPMD simulations with 16 beads the rate constant does not increase further and remains essentially unchanged within statistical uncertainties (see section S3.1 of the SI). More importantly, the mechanism is unchanged (only few trajectories do not follow a step-wise process) which allows to use only 8 beads for the free energy calculations and largely reduces their computational load.

To make sure that this effect does not depend on the size of the DNA model, we performed additional simulations with 5 base pairs (TTGAG sequence, corresponding to the addition of a thymine and a guanine at the beginning and at the end of the TGA sequence, respectively). No significant change was observed in the mechanism for these larger structures (see Figure S7 of the SI). Thus, we use the 3 base pairs model (with 8 beads for T-RPMD) for further discussions and US simulations.

In Figure 3, we show the free energy surfaces (FES) obtained for the isolated GC base pair and for 3BP-DNA system, as obtained from both classical and PIMD US simulations. The classical FES of the isolated dimer is almost symmetric with respect to the diagonal with a small deviation from the diagonal, of about 0.2 Å, which is in agreement with the corresponding direct dynamics results. When including NQEs, the barrier decreases, reflecting also in this case the results of the direct dynamics and in agreement with the FES reported on a simplified model by Perez et al.[15] Furthermore, the reaction pathway becomes fully symmetric as also found in direct dynamics simulations. In Table 2 we summarize the different free energy barriers associated with the process as obtained from DFTB-based

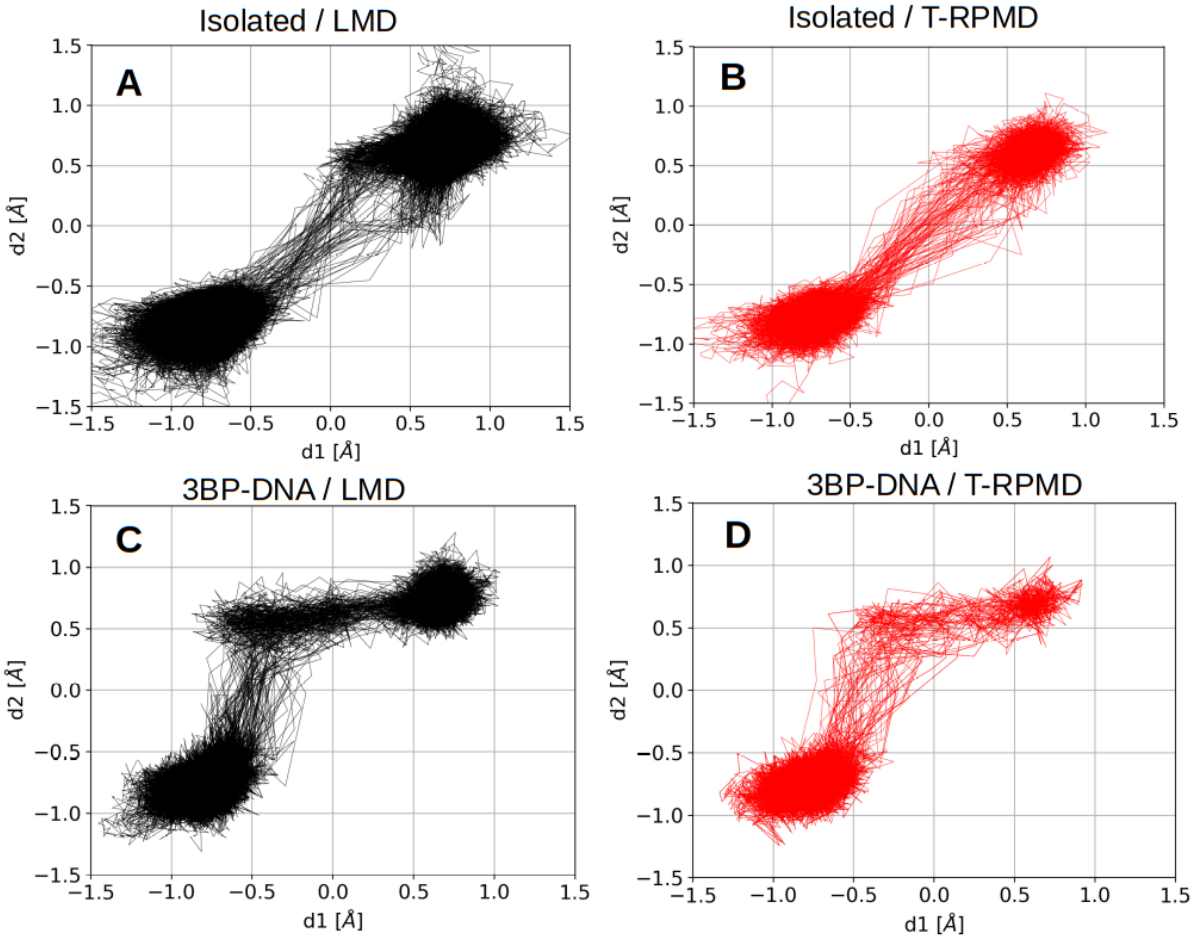
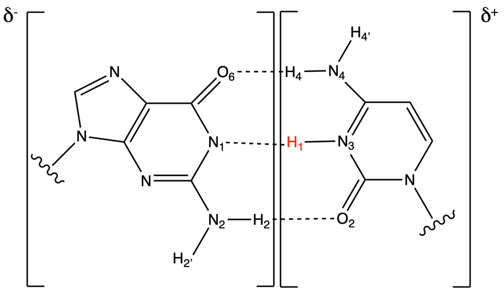


Figure 2: Trajectories projected on the d_1-d_2 collective variables plane, as obtained from DFTB-based direct dynamics simulations starting from the G*C* form: LMD and T-RPMD for the isolated base pair (panels A and B, respectively), LMD and T-RPMD for the DNA model environment (3BP-DNA, panels C and D).

US simulations, using both classical and Path Integral approaches. The inclusion of NQEs has an effect on both the free energy between GC and G*C* and on the associated barriers. Notably, the barrier associated with the reaction G*C* → GC decreases from 4.5 to 2.3 kcal/mol. However, the barrier associated with the GC → G*C* reaction remains relatively high, thus making the process unlikely without any particular source of activation.

When considering the DNA model environment, the FES profile changes dramatically: the tautomeric form becomes much less stable and the connection with the canonical form does not follow the diagonal (and consequently we cannot locate any saddle point in the FES analogous to the isolated system reaction). Notably, this is in agreement with the direct dynamics simulations, performed from the tautomeric to the canonical form. We now observe an intermediate form (shown in Scheme 1) in which only one proton has been transferred, corresponding to d_1 = -0.6 Å and d_2 = +0.6 Å. The d_1 coordinate, corresponding to the transferring proton exposed to the solvent, is the same as in the canonical form, while the d_2 coordinate is mostly in the tautomeric form. This new state is lower in free energy than the tautomeric form (14.4 vs 19.1 kcal/mol) and has a particular charge distribution character. In fact, while in the GC and G*C* forms the charge distribution is almost equally distributed between the two bases, in the intermediate, the guanine bears a negative charge (-0.8 e) and the cytosine almost a positive one (+0.6 e). Details on the charge calculations are reported in SI.



Scheme 1: Intermediate structure obtained from Umbrella Sampling simulations in the 3BP-DNA system.

When including NQEs, we do not see any major change in the free energy difference between GC and G*C* and the FES globally keeps a similar shape, but the intermediate is no more a clear minimum since the barrier connecting it to the canonical state is lost. We should note that the trajectories sample a region of the (d_1,d_2) space close to the intermediate minimum in the classical FES, but never reach it. This shows that the exact location of the intermediate as a free energy minimum is difficult to reach dynamically from the G*C*, but its presence has an important impact on the shape of the FES and consequently on the dynamics.

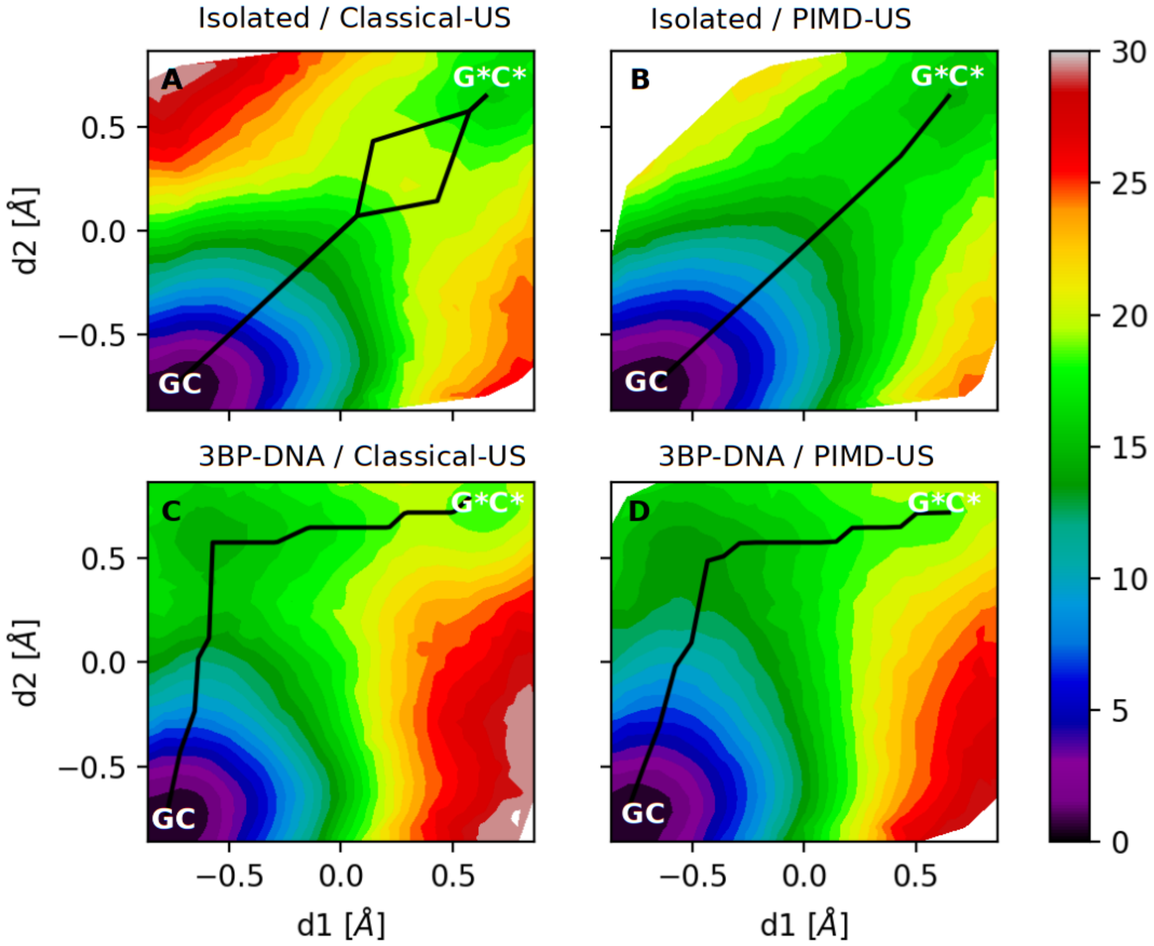


Figure 3: Free Energy Surfaces as a function of d_1 and d_2 collective variables as obtained from DFTB-based Umbrella Sampling (US) simulations: A) classical US of the isolated dimer; B) PIMD US of the isolated dimer; C) classical US of 3BP-DNA system; D) PIMD US of the 3BP-DNA system. Canonical (GC) and tautomeric (G*C*) states are indicated as well as the minimum free energy path connecting these two states as black solid line.

The decrease in the stability of the tautomeric form in the DNA model environment with

Table 2: Free energy differences (in kcal/mol) as obtained from US DFTB-based simulations for the isolated system and in the DNA model environment (3BP-DNA). The indices Cl and PI refer to classical and Path Integral Umbrella Sampling simulations, respectively, while ‡ denotes the free barrier to pass the saddle point of the given reaction. The "Intermediate" label refers to the locally stable state observed in Classical Umbrella Sampling in the 3BP-DNA model and represented in Scheme 1.

reaction	value	Isolated system	3BP-DNA
GC → G*C*	ΔF_{Cl}	15.0 ± 0.1	19.1 ± 0.3
GC → G*C*	ΔF_{PI}	14.1 ± 0.4	19.1 ± 0.2
GC → G*C*	$\Delta F^{\ddagger}_{Cl}$	19.5 ± 0.1	–
GC → G*C*	$\Delta F^{\ddagger}_{PI}$	16.4 ± 0.1	–
G*C* → GC	$\Delta F^{\ddagger}_{Cl}$	4.5 ± 0.1	–
G*C* → GC	$\Delta F^{\ddagger}_{PI}$	2.3 ± 0.4	–
GC → Intermediate	ΔF_{Cl}	–	14.4 ± 0.1

respect to the isolated system is also visible from the donor-acceptor distributions reported in Figure 4(A). Here we report data from LMD trajectories, the T-RPMD results are similar and reported in the Supporting Information, Figure S5. As shown in Figure 4(A), the donor-acceptor distance O_6N_4, corresponding to the external proton transfer, is shorter in the tautomeric structure form than in the canonical one. In the isolated system, this decrease in the distance is of about 0.04 Å, while in the DNA model it is close to 0.1 Å, showing that for this latter system, it is easier for the proton H_4 to move from the O_6 to N_4, causing the formation of the intermediate previously discussed.

An important finding of this study is that the DNA model environment has a crucial impact and modifies the DPT mechanism. By investigating the different trajectories we found that a key role is played by the surrounding water molecules (in the present simulations we included the crystallographic ones). More precisely, in Figure 4 (B), we show the distance distribution between O_6 and the two hydrogen atoms of the closest water molecule (labeled HW_1 and HW_2) for the three different forms: the canonical (green), the tautomeric (blue) and the intermediate forms (yellow). In the canonical form, the O_6 atom is strongly H-bonded to one water molecule (and of course to the H_4 atom of the cytosine base). In the tautomeric form, conversely, O_6 is covalently bound to the transferred H_4 atom and it is, therefore, less prone to form a hydrogen bond with the surrounding water. The nearby water molecule now interacts with O_6 via its two hydrogen atoms in a weaker way, as shown by the O_6-HW distributions. In the intermediate form, the water molecule moves back to the configuration where it makes a strong directional H-bond with O_6 (now H_4 is back to the cytosine as in the canonical form). This process is schematically shown in panels C and D of Figure 4 and a prototypical trajectory is reported as a movie file in the supporting material. In other words, the driving force that pushes the tautomeric form through the formation of the intermediate is the formation of an N–H bond and, from the point of view of the O_6 atom, the formation of two strong H-bonds: one with the cytosine base and one with the nearest water molecule.

Additional simulations of an isolated G*C* structure in which only few water molecules are included in the vicinity of the O–HN H-bond confirm this picture: the trajectories show a pathway which is similar to what is observed in the full DNA-model structure, as reported in the Supporting Information (Figure S7). These results, together with that of the 3BP-DNA model show the importance of micro-solvation on the reactivity. Notably, in agreement with the works by Tolosa et al.[45] and Gheorghiu et al.[4] with different methods and approaches, micro-solvation is crucial to open a new reaction pathway connecting GC with G*C* via an intermediate. An important finding of the present study is that this intermediate is partially destabilized by NQEs.

Summarizing, we report an exhaustive study of how NQEs and the environment can affect the mechanism of DPT in the GC base pair, using a combination of direct dynamics simulations, to characterize the spontaneous decay of the tautomeric form, and Umbrella sampling simulations to obtain the free energy surface of this reaction.

For the isolated dimer, the mechanism is concerted though slightly asynchronous and the effect of NQEs is to accelerate the reaction by approximately a factor of 30, making the mechanism fully synchronous. Indeed, while in LMD simulations the minimum free energy pathway passes about 0.2 Å away from the diagonal of d_1-d_2 plot, it moves to follow the

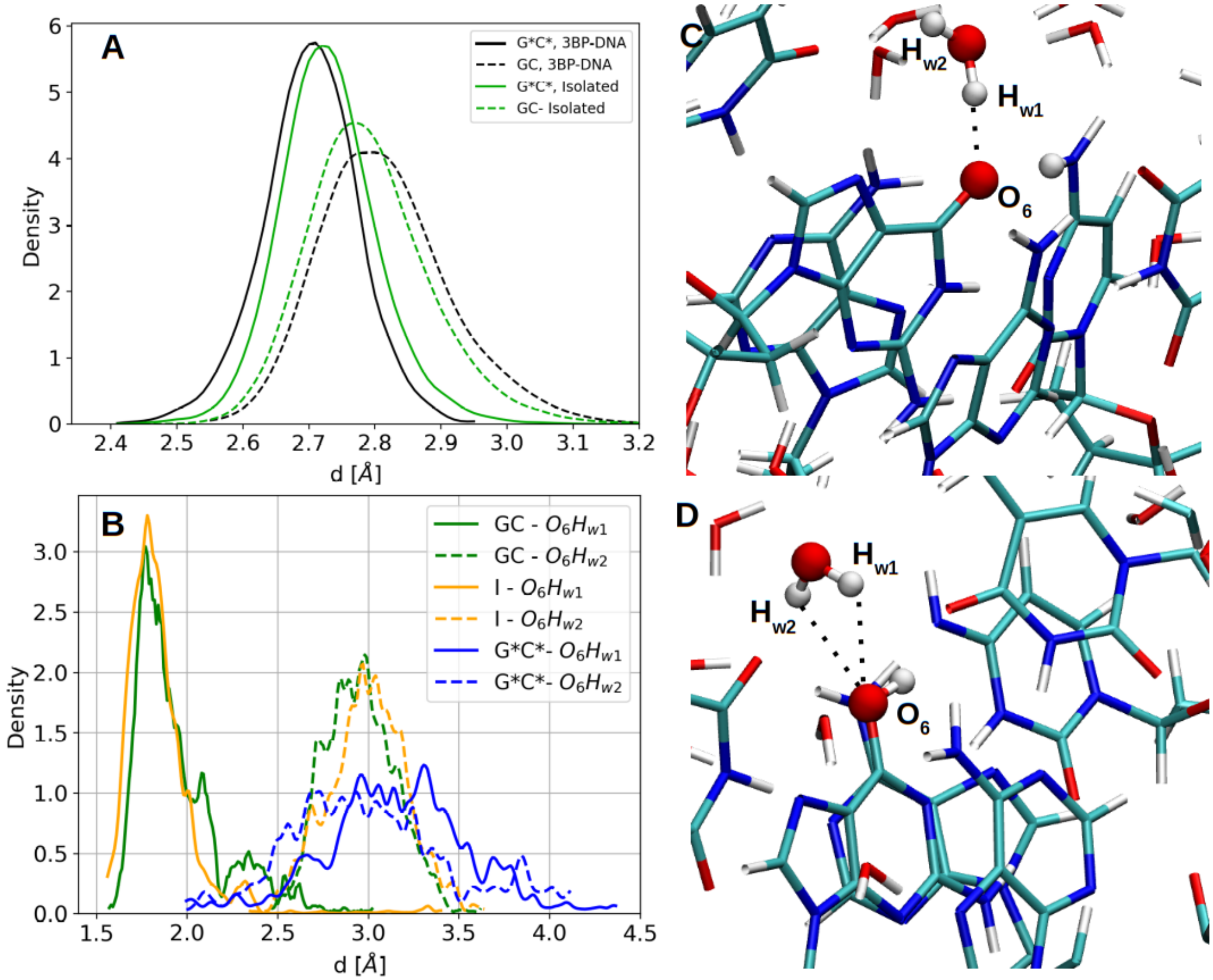


Figure 4: Panel A: O–N distance distributions in the isolated system (green) and in the 3BP-DNA (black) as obtained for the canonical (dashed line) and the tautomeric (continuous line) forms from DFTB-based simulations. Panel B: Distance distributions between the O_6 and hydrogen atoms of the nearest water molecule (H_{w1} and H_{w2}) obtained from DFTB-based classical US simulations: canonical (green), tautomeric (blue) and intermediate (yellow) forms. Panels C and D: two prototypical snapshots of the canonical (C) and tautomeric (D) forms where the atoms involved in DPT reaction are highlighted as balls.

diagonal almost exactly when NQEs are included. When the environment is taken into account in the 3BP-DNA structure, NQEs still speed up the DPT process of about 10 times compared to the classical LMD results. The free energy landscape is also modified such as the intermediate form ceases to be a local minimum, making the proton transfer essentially barrierless. Therefore, our study shows the importance of the environment which completely changes the mechanism of this process, which becomes a step-wise reaction. In particular, we show the importance of the role of the surrounding water molecules, that stabilize an intermediate structure with opposite partial charges on each base pair.

Our results clearly show that further computational studies must include the environment at least by considering more than one base pair and (crystallographic) water molecules. This will be important to investigate if modifications in the base pairs, for example through methylation which may result from carcinogenic agents,[46,47] have an effect on the stability of the tautomeric and/or intermediate forms. Finally, the step-wise mechanism reported for GC tautomerism could be important to unravel the biochemistry of GC-rich DNA regions which are associated with the so-called CpG islands[48] and related to the promoter region of the genome.[49,50]

Acknowledgement

We thank ED 388 from Sorbonne Université for doctoral studentship and Dr. Jessica Bowles for careful reading of the manuscript.

Supporting Information Available

In the supporting information we report: (1) Details on DFTB benchmarking; (2) Details of reaction dynamics and Umbrella sampling simulations; (3) Additional results in terms of population decays, 16-beads results, distance distributions and direct dynamics trajectories with 5 base pairs, smaller time step and crystallographic water molecules around the isolated base pair. Details on the charge calculation procedure are also provided with the corresponding results.

We also provide as supporting material a movie file showing a prototypical reaction.

References

(1) Modrich, P. DNA mismatch correction. *Ann. Rev. Biochem.* **1987**, *56*, 435–466.

(2) Lowdin, P.-O. Proton tunneling in DNA and its biological implications. *Rev. Mod. Phys.* **1963**, *35*, 742–732.

(3) Cerón-Carrasco, J. P.; Requena, A.; Michaux, C.; Perpéte, E. A.; Jacquemin, D. Effects of hydration on the proton transfer mechanism in the adenine-thymine base pair. *J. Phys. Chem. A.* **2009**, *113*, 7892–7898.

(4) Gheorghiu, A.; Coveney, P. V.; Arabi, A. A. The influence of base pair tautomerism on single point mutations in aqueous DNA. *Interface Focus* **2020**, *10*, 20190120.

(5) Cerón-Carrasco, J. P.; Requena, A.; Zuniga, J.; Michaux, C.; Perpéte, E. A.; Jacquemin, D. Intermolecular Proton Transfer in Microhydrated Guanine-Cytosine Base Pairs: a New Mechanism for Spontaneous mutation in DNA. *J. Phys. Chem. A.* **2009**, *113*, 10549–10556.

(6) Galtier, N.; Piganeau, G.; Mouchiroud, D.; Duret, L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* **2001**, *159*, 907–911.

(7) Catalán, J.; del Valle, J. C.; Kasha, M. Resolution of concerted versus sequential mechanisms in photo-induced double-proton transfer reaction in 7-azaindole H-bonded dimer. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 8338–8343.

(8) Pohl, R.; Socha, O.; Slavicek, P.; Sala, M.; Hodgkinson, P.; Dracinsky, M. Proton transfer in guanine-cytosine base pair analogue studied by NMR spectroscopy and PIMD simulations. *Faraday Discuss.* **2018**, *212*, 331–344.

(9) Li, P.; Rangadurai, A.; Al-Hashimi, H.; Hammes-Schiffer, S. Environmental effects on guanine-thymine mispair tautomerization explored with quantum mechani-

cal/molecular mechanical free energy simulations. *J. Am. Chem. Soc.* **2020**, *142*, 11183–11191.

(10) Soler-Polo, D.; Mendieta-Moreno, J. I.; Trabada, D. G.; Mendieta, J.; Ortega, J. Proton Transfer in Guanine-Cytosine Base Pairs in B-DNA. *J. Chem. Theory Comput.* **2019**, *15*, 6984–6991.

(11) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(12) Fang, W.; Chen, J.; Feng, Y.; Li, X.-Z.; Michaelides, A. The quantum nature of hydrogen. *Int. Rev. Phys. Chem.* **2019**, *38*, 35–61.

(13) Litman, Y.; Richardson, J. O.; Kumagai, T.; Rossi, M. Elucidating the Nuclear Quantum Dynamics of Intramolecular Double Hydrogen Transfer in Porphycene. *J. Am. Chem. Soc.* **2019**, *141*, 2526–2534.

(14) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. On the Quantum Nature of the Shared Proton in Hydrogen Bonds. *Science* **1997**, *275*, 817–820.

(15) Pérez, A.; Tuckerman, M. E.; Hjalmarson, H. P.; von Lilienfeld, O. A. Enol Tautomers of WatsonCrick Base Pair Models Are Metastable Because of Nuclear Quantum Effects. *J. Am. Chem. Soc.* **2010**, *132*, 11510–11515.

(16) Marx, D.; Parrinello, M. Ab initio path integral molecular dynamics: Basic ideas. *J. Chem. Phys.* **1996**, *104*, 4077–4082.

(17) Slocombe, L.; Al-Khalili, J.; Scacchi, M. Quantum and classical effects in DNA point mutations: Watson-Crick tautomerism in AT and GC base pairs. *Phys. Chem. Chem. Phys.* **2021**, *23*, 4141.

(18) Slocombe, L.; Al-Khalili, J.; Scacchi, M. An Open Quantum Systems approach to proton tunnelling in DNA. *Commun. Phys.* **2022**, *23*, 109.

(19) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(20) Nieman, R.; Spezia, R.; Jayee, B.; Minton, T.; Hase, W. L.; Guo, H. Exploring Reactivity and Product Formation in N(4S) Collisions with Pristine and Defected Graphene with Direct Dynamics Simulations. *J. Chem. Phys.* **2020**, *153*, 184702.

(21) Malik, A.; Spezia, R.; Hase, W. L. Unimolecular Fragmentation Properties of Thermometer Ions from Chemical Dynamics Simulations. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 169–179.

(22) Young, T. A.; Johnston-Wood, T.; Zhang, H.; Duarte, F. Reaction dynamics of Diels-Alder reactions from machine learned potentials. *Phys. Chem. Chem. Phys.* **2022**, *24*, 20820.

(23) Narayana, N.; Ginell, S. L.; Russu, I. M.; Berman, H. M. Crystal and molecular structure of a DNA fragment: d(CGTGAATTCACG). *Biochem.* **1991**, *30*, 4449–4455.

(24) Cerón-Carrasco, J. P.; Jacquemin, D. DNA spontaneous mutation and its role in the evolution of GC-content: assessing the impact of the genetic sequence. *Phys. Chem. Chem. Phys.* **2015**, *17*, 7754–7760.

(25) Ivanov, S. D.; Grant, I. M.; Marx, D. Quantum free energy landscapes from ab initio path integral metadynamics: Double proton transfer in the formic acid dimer is concerted but not correlated. *J. Chem. Phys.* **2015**, *143*, 124304.

(26) Xiao, S.; Wang, L.; Liu, Y.; Lin, X.; Liang, H. Theoretical investigation of the proton transfer mechanism in guanine-cytosine and adenine-thymine base pairs. *J. Chem. Phys.* **2012**, *137*, 195101.

(27) Craig, I. R.; Manolopoulos, D. E. Quantum statistics and classical mechanics: Real time correlation functions from ring polymer molecular dynamics. *J. Chem. Phys.* **2004**, *121*, 3368–3373.

(28) Craig, I. R.; Manolopoulos, D. E. A refined ring polymer molecular dynamics theory of chemical reaction rates. *J. Chem. Phys.* **2005**, *123*, 034102.

(29) Fang, W.; Chen, J.; Rossi, M.; Feng, Y.; Li, X.-Z.; Michaelides, A. Inverse Temperature Dependence of Nuclear Quantum Effects in DNA Base Pairs. *J. Phys. Chem. Lett.* **2016**, *7*, 2125–2131.

(30) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(31) Bondar, A.-N.; Fischer, S.; Smith, J. C.; Elstner, M.; Suhai, S. Key Role of Electrostatic Interactions in Bacteriorhodopsin Proton Transfer. *J. Am. Chem. Soc.* **2004**, *126*, 14668–14677.

(32) Hourahine, B. et al. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **2020**, *152*, 124101.

(33) Brovarets, O.; Hovorun, D. Can tautomerization of the A·T Watson–Crick base pair via double proton transfer provoke point mutations during DNA replication? A comprehensive QM and QTAIM analysis. *J. Biomol. Struct. Dyn.* **2014**, *32*, 127–154.

(34) Yanai, T.; Tew, D.; Handy, N. A new hybrid exchange-correlation functional using the Coulomb-Attenuating Method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(35) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comp. Chem.* **2011**, *32*, 1456–1465.

(36) Ceriotti, M.; Parrinello, M.; Markland, T. E.; Manolopoulos, D. E. Efficient stochastic thermostatting of path integral molecular dynamics. *J. Chem. Phys.* **2010**, *133*, 124104.

(37) Hele, T. J.; Suleimanov, Y. V. Should thermostatted ring polymer molecular dynamics be used to calculate thermal reaction rates? *J. Chem. Phys.* **2015**, *143*, 074107.

(38) Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(39) Grossfield, A. WHAM: the weighted histogram analysis method. *http://membrane.urmc.rochester.edu/wordpress/?page_id=126* **Version 2.10**,

(40) Hinsen, K.; Roux, B. Potential of mean force and reaction rates for proton transfer in acetylacetone. *J. Chem. Phys.* **1997**, *106*, 3567–3577.

(41) Suleimanov, Y. V.; Collepardo-Guevara, R.; Manolopoulos, D. E. Bimolecular reaction rates from ring polymer molecular dynamics: Application to H + CH4→ H2 + CH3. *J. Chem. Phys.* **2011**, *134*, 044131.

(42) Angiolari, F.; Huppert, S.; Spezia, R. Quantum versus Classical Unimolecular Fragmentation Rate Constants and Activation Energies at Finite Temperature from Direct Dynamics Simulations. *Phys. Chem. Chem. Phys.* **2022**, *24*, 29357–29370.

(43) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comp. Phys. Commun.* **2014**, *185*, 604–613.

(44) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comp. Phys. Commun.* **2009**, *180*, 1961–1972.

(45) Tolosa, S.; Sanson, J.; Hidalgo, A. Mechanisms for guanine-cytosine tautomeric equilibrium in solution via steered molecular dynamics simulations. *J. Mol. Liq.* **2018**, *251*, 308–316.

(46) Peterson, L. A.; Hecht, S. S. O6-Methylguanine is a critical determinant of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone tumorigenesis in A/J mouse lung. *Cancer Res.* **1991**, *51*, 5557–5564.

(47) Hecht, S. S. DNA adduct formation from tobacco-specific N-nitrosamines. *Mut. Res.-Fund. Mol. M.* **1999**, *424*, 127–142.

(48) Aïssani, B.; Bernardi, G. CpG islands, genes and isochores in the genomes of vertebrates. *Gene* **1991**, *106*, 185–195.

(49) Saxonov, S.; Berg, P.; Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1412–1417.

(50) Deaton, A. M.; Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **2011**, *25*, 1010–1022.