



Replacing thymine with a strongly pairing fifth Base: A combined quantum mechanics and molecular dynamics study



Mohit Chawla^{a,*}, Suresh Gorle^b, Abdul Rajjak Shaikh^a, Romina Oliva^{c,*}, Luigi Cavallo^{a,*}

^a King Abdullah University of Science and Technology (KAUST), Physical Sciences and Engineering Division, Kaust Catalysis Center, Thuwal 23955-6900, Saudi Arabia

^b Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555, USA

^c Department of Sciences and Technologies, University Parthenope of Naples, Centro Direzionale Isola C4, I-80143 Naples, Italy

ARTICLE INFO

Article history:

Received 4 December 2020

Received in revised form 6 February 2021

Accepted 9 February 2021

Available online 23 February 2021

Keywords:

Nucleic acids

Quantum chemical studies

Molecular dynamic studies

Non-natural base pairs

Synthetic biology

Modified bases

ABSTRACT

The non-natural ethynylmethylpyridone C-nucleoside (W), a thymidine (T) analogue that can be incorporated in oligonucleotides by automated synthesis, has recently been reported to form a high fidelity base pair with adenosine (A) and to be well accommodated in B-DNA duplexes. The enhanced binding affinity for A of W, as compared to T, makes it an ideal modification for biotechnological applications, such as efficient probe hybridization for the parallel detection of multiple DNA strands. In order to complement the experimental study and rationalize the impact of the non-natural W nucleoside on the structure, stability and dynamics of DNA structures, we performed quantum mechanics (QM) calculations along with molecular dynamics (MD) simulations. Consistently with the experimental study, our QM calculations show that the A:W base pair has an increased stability as compared to the natural A:T pair, due to an additional CH- π interaction. Furthermore, we show that mispairing between W and guanine (G) causes a distortion in the planarity of the base pair, thus explaining the destabilization of DNA duplexes featuring a G:W pair. MD simulations show that incorporation of single or multiple consecutive A:W pairs in DNA duplexes causes minor changes to the intra- and inter-base geometrical parameters, while a moderate widening/shrinking of the major/minor groove of the duplexes is observed. QM calculations applied to selected stacks from the MD simulations also show an increased stacking energy for W, over T, with the neighboring bases.

© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Molecular recognition in the double helical structure of DNA follows H-bonding complementarity, where the hydrogen bond donors on one nucleobase pair with the hydrogen bond acceptors on the opposite base. In natural DNA structures, the Watson-Crick pairing occurs within the two canonical base pairs, A:T and G:C, with the latter featuring a higher stability, due to the formation of one more strong H-bond. Besides the canonical A/T(U)/G/C bases, however, a number of natural modified bases have been reported, in DNA and especially in RNA molecules [1,2].

In addition, during the past three decades, a number of synthetic nucleotide analogues have been incorporated in DNA and RNA molecules for a range of applications, including the probing of biological interactions and the expansion of the genetic alphabet

[3–10]. An active area of research in bioengineering and therapeutic applications is the synthesis of nucleobase analogues that possess enhanced binding affinity when paired to their complementary bases. They would allow, for example, an efficient probe hybridization, independently of the complementary sequence, thus facilitating the parallel detection of multiple DNA strands [11,12]. Examples of such modifications include expanded bases [13,14], clamp-like base derivatives [15] as well as bases with ankynyl functionalization [16,17].

Regarding this last class of modified bases, in 2013 Minuth and Richert reported on the C-nucleoside 6-ethynylpyridone, abbreviated as E (Fig. 1), a T analogue where an ethynyl group replaced the carbonyl oxygen at the C2 position. This functional group can potentially give a CH- π interaction with the C2-H atom on A and is also expected to strengthen the stacking interactions with the neighboring bases in the duplex structure [18]. Experimental results evidenced that E indeed H-bonds to A more strongly than T. In fact, UV-melting experiments confirmed that the replacement of a T:A pair by an E:A pair increased the melting temperature of a

* Corresponding authors.

E-mail addresses: Mohit.chawla@kaust.edu.sa (M. Chawla), romina.oliva@uni-parthenope.it (R. Oliva), luigi.cavallo@kaust.edu.sa (L. Cavallo).

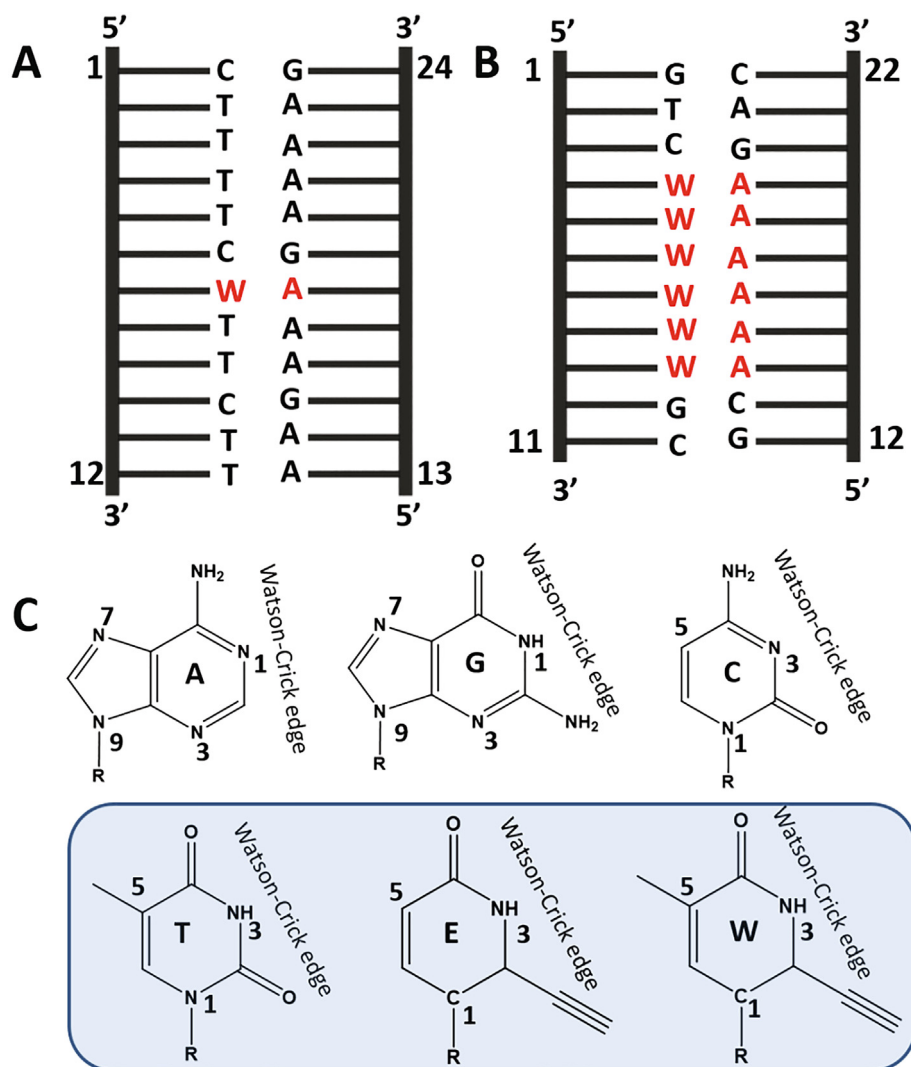


Fig. 1. A/B) Sequence of the 12-mer (1AW) and 11-mer (6AW) double-stranded B-DNA oligomers, from Ref. [12], featuring 1 or 6 central A:W pairs, respectively; the A:W pairs are highlighted in red. C) Chemical structures of the four natural nucleobases in DNA (A, G, C, T) and of the non-natural E and W bases. In order to emphasize their similarity, T, E and W are highlighted in a blue shaded box. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

12-mer DNA duplex by an extent comparable to that observed for the substitution of T:A with G:C (2.6 vs 2.8 °C) [18]. This confirmed that the A:E base pair is almost as strong as the G:C pair. A subsequent theoretical study, based on quantum chemistry and focusing on H-bonding, indicated that the energetic stability of the A:E pair is intermediate between those of a A:T and a G:C base pair [19]. Later on, a dispersion corrected DFT study addressed the impact of stacking interactions between E and the neighboring bases, showing a strengthening in the stacking of the A:E base pair as compared to the canonical base pairs [20]. Taken together, results of such calculations explain the increased stability of duplexes incorporating the non-natural E base facing a natural A. However, E has to be incorporated into DNA by a manual coupling after strand phosphorylation on solid support, resulting in low yield. Furthermore, it lacks the methyl group at the C5 position, which makes it difficult to be recognized by proteins on the major groove, including the crucial repair enzymes [18,21].

To circumvent this problem, in a recent study, Walter and Richert reported the synthesis of the ethynylmethylpyridone C-nucleoside, abbreviated as W, a novel T analogue that features

an ethynyl group at C2 while preserving the methyl group at C5 (see Fig. 1). The modified W could be incorporated in high yield by automated synthesis in several oligonucleotides, where it caused significant increase in the melting temperatures over the corresponding A:T containing duplexes [12]. In addition, it exhibited high pairing fidelity, with a single G:W mismatch causing a dramatic melting point depression. On these grounds, W has been proposed as an ideal “fifth base”, to be used to target weakly pairing A-rich sequences in biological applications [12].

To complement the reported experimental study and to rationalize the impact of the non-natural W base on the stability and dynamics of nucleic acids structure, we have performed quantum mechanics (QM) calculations along with classical molecular dynamics (MD) simulations, both having been proven to be effective in the study of base pairing and stacking interactions in nucleic acids [12,22–31]. QM calculations have initially been focused on the geometry and energetics of the H-bonding for the non-natural A:W base pair, which have then been compared to those of the natural A:T and G:C Watson-Crick pairs. Next, we have investigated the geometry and energetics of possible H-bonded mismatched pairs between W and the natural G/C/T bases.

Further, to investigate the structure and stability of A:W in the context of real DNA structures and the possible structural perturbation induced due to the incorporation of an ethynyl group on C2 atom, we have also carried out MD simulations of two of the duplexes as investigated in [12]. In particular, we have simulated a 12-mer DNA duplex featuring a single central A:W base pair (abbreviated by us as 1AW-DNA from here on), and a 11-mer duplex including six central A:W pairs (abbreviated by us as 6AW-DNA; see Fig. 1). Both these duplexes were experimentally shown to exhibit increased melting temperatures as compared to the corresponding A:T containing duplexes, by 4.4 and 17.5 °C, respectively. Finally, snapshots from the MD simulations have been extracted to investigate by QM techniques the impact of W on the stacking interactions with adjacent bases in the duplex.

As a result of these analyses, we can now dissect and explain the effect of the W modification on the DNA geometry, stability, and dynamics.

2. Materials and methods

2.1. Quantum mechanics calculations

All systems were investigated in the canonical *cis* Watson Crick geometry. They correspond to the four non-natural base pairs: A:W, G:W, C:W, T:W, and the corresponding natural pairs: A:T, G:T, C:T, T:T. For all the systems described above, the base pairs were truncated at the C1' atom of the ribose. This is the standard approach used in literature [23,26,32–38].

Geometry optimizations were performed within a DFT approach, based on the hybrid B3LYP functional as implemented in the Gaussian09 package [39–41]. The correlation-consistent polarized valence triple- ζ cc-pVTZ basis set [42], was used for all the geometry optimizations in gas phase as well as in water, modeled with the C-PCM continuum solvation model [43]. Since dispersion interactions might contribute differently to the stability of the base pairs under study, we also added an empirical dispersion term, with the Becke-Johnson damping scheme, to the electronic energy [44]. Interaction energies were calculated on the B3LYP-D3BJ/cc-pVTZ optimized geometries at the second order Møller-Plesset (MP2) [45] level of theory using the augmented aug-cc-pVTZ basis set. For these calculations, we used the RIMP2 [46] method as implemented in Turbomole 6.1 package, with water modeled with the continuum solvation model COSMO [43]. All the interaction energies were corrected for the basis set superposition error (BSSE) using the counterpoise procedure proposed by Boys and Bernardi [47]. Thus, the binding energy E_{Bind} is calculated as in Eq. (1):

$$E_{\text{Bind}} = E_{\text{Complex}} - (E_{\text{M1}} + E_{\text{M2}}) + \text{BSSE}; \quad (1)$$

where E_{Complex} is the electronic energies of the optimized M1:M2 base pair, and E_{M1} and E_{M2} are the electronic energy of the isolated M1 and M2 bases, and BSSE is the basis set superposition error. Within this approach the deformation energy, which is the energy required to deform the bases from the isolated geometry to the geometry they have in the base pair, is included in our calculations. This is a rather standard approach used in this kind of calculations [22,27,48–52]. In the present study, we also derived the interaction energies in water, which were calculated using the same recipe as suggested by Spomer and coworkers [48,53].

To have an immediate and intuitive understanding of the impact of a specific modification, we introduced the modification energy, E_{Mod} , defined as the energy difference between the binding energy of the modified/non-natural pair (for instance W:A) and of the corresponding base pair where the non-natural base pair was substituted with the natural one [23], as shown in Eq. (2).

$$E_{\text{Mod}} = E_{\text{Bind}}(\text{modified}) - E_{\text{Bind}}(\text{unmodified}). \quad (2)$$

Within this definition, negative and positive E_{Mod} values indicate increase or decrease in the stability of a specific base pair, respectively, as compared to the reference natural base pair system.

We evaluated the base-base stacking interaction energy as described in Eq. (1), using the RIMP2 method with aug-cc-pVTZ basis set and BSSE correction. Also for these calculations, the sugar-phosphate backbone was removed and the bases were truncated at the C1' position with a methyl group.

In order to evaluate the accuracy of the calculated RIMP2 single point energies, we have also performed single point energy calculations using the DLPNO-CCSD(T) method of Neese and co-workers [54–56] for all the base pairs under investigation as implemented in ORCA 4.0 [57]. Tighter than the default “TightPNO” DLPNO settings (TCutPairs = 10^{-5} , TCutPNO = 10^{-7} , and TCutMKN = 10^{-3}) were used [58]. The triple and quadruple- ζ correlation consistent basis sets of Dunning augmented with diffuse functions were used in the present work to describe hydrogen, carbon, nitrogen and oxygen atoms [42]. The correlation fitting basis sets aug-cc-pVQZ/C developed by Hättig and co-workers [59], necessary for the resolution of identity approximation as a part of DLPNO scheme, were used. All aug-cc-pVQZ/C basis sets were used as implemented in ORCA 4.0 suite of programs [57].

To account for the basis set incompleteness effects, we applied the extrapolation schemes for Hartree-Fock and DLPNO-CCSD(T) correlation energies proposed by Helgaker and co-workers [60–62], see Eqs. (3) and (4). For two adjacent triple and quadruple- ζ basis sets:

$$E_{\text{HF}}^n = E_{\text{HF}}^\infty + \alpha e^{-1.63n} \quad (3)$$

$$E_{\text{corl}}^n = E_{\text{corl}}^\infty + \beta n^{-3} \quad (4)$$

where $n = 3$ and 4 for triple and quadruple- ζ basis sets; $E_{\text{HF}}^\infty/E_{\text{corl}}^\infty$ are the Hartree-Fock and correlation energies at the complete basis set (CBS) limit; α/β are parameters to be obtained from a system of the two equations.

2.2. Molecular dynamics simulations

2.2.1. Model building

The DNA sequences, 1AW-DNA and 6AW-DNA (see Fig. 1), used in the MD simulations correspond to the duplexes 19 and 18, respectively, from the experimental study [12]. Initially, the corresponding wild type B-DNA duplexes, featuring A:T in place of the A:W pairs, were built by using the Make-NA server (<http://structure.usc.edu/make-na/server.html>), then the coordinates of the central nucleobases (1 in 1AW and 6 in 6AW) were replaced with those of the geometrically optimized A:W pair by using VMD [63].

2.2.2. Parametrization of W

The geometrically optimized coordinates of the W nucleoside structures obtained from QM calculations (see above) were used to derive the structural force field parameters (bonds, angles and dihedrals) of the nucleobase portion for the non-natural W nucleotide. The force constants were assigned taking into account analogous values in standard nucleobase fragments, included in the Amber bsc1 force field library [64]. The electrostatic potential surface was generated by the Merz-Kollman method at the HF/6-31G(d) level of theory, followed by multi configurational two-stage RESP fitting using the RED IV program [65,66], by using the standard protocol proposed by Cornell et al. [67]. All the parameters corresponding to the deoxyribose sugar moiety connecting to the non-natural bases and the phosphate group were derived from the Amber bsc1 force field library values [64].

2.2.3. Simulation protocol

All the MD simulations were performed with NAMD [68] in a water box with $50 \text{ \AA} \times 52 \text{ \AA} \times 66 \text{ \AA}$ dimensions solvated using the TIP3P water model [69] and simulating standard biological conditions by considering a 150 mM NaCl concentration with additional 22Na^+ ions to neutralize the system. Electrostatic interactions were treated using the particle mesh Ewald method and a non-bonded cutoff of 1.4 nm was used for the Lennard-Jones potential. The simulations were run under NPT conditions (298 K and 1 bar) using the Berendsen algorithm [70] to control temperature and pressure, with a coupling constant of 5 ps for both parameters, a benchmark protocol by the ABC consortium (<https://bisi.ibcp.fr/ABC/Protocol.html>). Periodic boundary conditions were employed and the SHAKE algorithm [71] was used to constrain all bonds involving hydrogens. An integration time step of 2 fs was used and the coordinates were saved every 5 ps for further analyses. In both systems the water molecules were relaxed by performing energy minimization and followed by 500 ps of MD in NVT ensemble at 298 K, restraining the atomic positions with a weak harmonic potential. The systems were heated up gradually to 298 K in a six-step process, starting from 50 K to 298 K. After heating, the systems were simulated under NPT standard conditions, without restraints, for 0.5 μs (500 ns).

2.2.4. Post-simulation analyses

The analyses of the MD trajectories were done by the VMD [63], Pymol [72] and GnuPlot [73]. For the interaction energies calculation, 800 MD frames each were extracted from the trajectories of 1AW-DNA and 1AT-DNA, every 0.5 ns, between 100 ns and 500 ns of simulation time. Interaction energies for the central A:W and A:T base pairs were calculated with the NAMDEnergy plugin of VMD, as the difference between the total energy of the nucleobase pair and the sum of total energies of the individual nucleobases, and averaged over the 800 frames. Finally, 5 frames every 100 ns were extracted from the MD simulations for each system, to calculate the stacking interaction energies of the non-natural W base, as compared to A, with the adjacent neighboring bases by a QM approach (see QM calculations). The glycosidic bond preference was assessed by measuring the dihedrals between atoms O4'-C1'-N9-C4 (for A), O4'-C1'-N1-C2 (for T), and O4'-C1'-C1-C2 (for W). The Curves+ program [74] was employed for the calculation of helical parameters.

3. Results and discussion

3.1. Optimal geometries and stability of the W-containing base pairs

Eight base pair combinations have been investigated in their canonical *cis* Watson-Crick geometry (abbreviated as cWW from here on [75]). They include the base pairs given by W with the four natural DNA bases, A, G, C, T, all previously inserted in the middle of a model DNA duplex and experimentally characterized [12], and the corresponding unmodified base pairs, A:T, G:T, C:T and T:T, as reference systems.

For all the above base pairs, we calculated the optimal geometry and interaction energy both in the gas phase and in water. To have an immediate understanding of the comparative stability of a natural vs. a W-containing base pair, we calculated the “modification” energy, E_{mod} (see Methods). By definition, a negative E_{mod} value implies that a modified pair is more stable than the corresponding natural pair (and *vice versa*). All the calculated energies are reported in Table 1. The obtained optimal geometries in the gas phase are shown in Fig. 2, with the corresponding H-bond distances calculated in the gas phase and in water reported.

3.1.1. A:W cWW base pair

The optimized geometry of A:W presents two strong H-bonds, along with an additional CH- π interaction between the C2-H atom of A and the ethynyl group of the W base. The two strong hydrogen bonds are identical to those of A:T, with a slight elongation of the N1(A)-N3(W) H-bond and a slight compression of the N6(A)-O4(W) H-bond, by $\approx 0.09 \text{ \AA}$ each. The C1'-C1' distance in the gas-phase optimized geometry is 10.88 \AA , that is extremely similar to that of the canonical G:C/A:T cWW pairs (10.69/10.53 \AA). This means that this base pair is perfectly isosteric with the natural base pairs found in DNA. Optimization of the A:W pair in water resulted in a geometry consistent with the gas phase one, with differences in the calculated H-bond distances and the C1'-C1' distances within 0.02 \AA .

Focusing on energetics, the non-natural A:W pair is more stable than A:T (with E_{mod} of -2.20 kcal/mol and -1.45 kcal/mol in the gas phase and in water, respectively), which is in line with the experimental finding that replacement of T with W increases the stability of a DNA duplex [12]. It is to be remarked that the electrostatic component of the interaction energy gets attenuated by solvent screening in water, resulting in significantly lower interaction energies for the A:W/A:T base pairs as compared to the gas phase calculations.

To shed light on the enhanced stability of the A:W pair relatively to A:T, we compared the electron densities of both the systems. Differences in the electron density between the W and T bases and the A:W and A:T base pairs in the gas phase are plotted in Fig. 3A, B. From the figure, it is clear that a depletion in electron density is observed around the O4 atom of W, which makes it a poorer H-bond acceptor. The electron density is instead increased around the N6 atom of the A base and decreased around its bound hydrogen, thus making N6(A) a better H-bond donor. These opposite effects result overall in a stronger H-bond between N6(A) and O4(W), as compared to the corresponding H-bond in the unmodified A:T pair. On the contrary, the H-bond between N1(A) and N3(W) is weaker than the corresponding one in the unmodified base pair, as an increase in the electron density is observed around the hydrogen on N3(W), making it a poorer H-bond donor, while the electron density around N1(A) is unchanged in the two base pairs. In agreement with the above analysis, the H-bond distances of the optimized structures (see Fig. 2) show a respective shrinking and elongation of the N6(A)-O4(W) and N1(A)-N3(W) H-bonds in A:W as compared to A:T. Therefore, the balance of these effects result in a similar stability of the two base pairs in terms of H-bonding.

The short distance between the C2-H group on A and the ethynyl group on W suggests that the increased stability of the A:W pair could be explained by an attractive CH- π type interaction between these groups. To shed light on this point, we performed a Non Covalent Interaction (NCI) analysis, using the approach developed by Yang and coworkers [76,77] (Fig. 3C, Fig. S1). As supposed, in addition to the expected attractive interactions corresponding to the above H-bonds (blue patches), the NCI isosurface of the A:W base pair shows a similarly attractive interaction between C2-H on A and the ethynyl group on W.

3.1.2. T:W cWW base pair

The optimized geometry of T:W in the gas phase is planar and stabilized by two strong H-bonds, similarly to the natural T:T pair. Its C1'-C1' distance is 8.62 \AA , which compares with 8.55 \AA in T:T. Not surprisingly for the pairing between two 6-membered rings, this distance is significantly shorter (by $\approx 2 \text{ \AA}$) than that of the canonical A:T/G:C pairs. In water optimization results in a geometry which is perfectly consistent with the gas phase optimized one, with differences in the calculated H-bond and C1'-C1' distances within 0.08 \AA . Regarding energy, it is more stable than T:T, with

Table 1

Interaction energies, E_{int} , and binding energies, E_{bind} , for the cWW A:T and G:C base pairs and the modified pairs involving W. All energies are reported in kcal/mol. E_{Mod} is the difference between the binding energy of the modified base pair and of the reference natural pair. Negative and positive values of E_{Mod} indicate that the modified base pair is more stable or less stable than the reference pair. Energies calculated at the RIMP2/aug-cc-pVTZ level on B3LYP-D3BJ/cc-pVTZ optimized geometries and E_{bind} and E_{Mod} values calculated by CCSD(T)-DLPNO/CBS in the gas phase are reported.

System	E_{int} gas (MP2)	E_{def} gas (MP2)	E_{bind} gas (MP2)	E_{bind} gas (DLPNO)	E_{Mod} gas (MP2)	E_{Mod} gas (DLPNO)	E_{bind} water (MP2)	E_{Mod} water (MP2)
G:C	-30.74	2.75	-27.98	-28.08	-	-	-12.49	-
A:T	-16.35	1.43	-14.92	-15.01	-	-	-7.89	-
A:W	-18.50	1.38	-17.12	-16.41	-2.20	-1.40	-9.34	-1.45
C:T	-15.30	1.90	-13.40	-13.45	-	-	-6.32	-
C:W	-19.73	1.69	-18.04	-17.66	-4.64	-4.21	-8.48	-2.16
T:T	-13.39	0.98	-12.41	-12.53	-	-	-7.01	-
T:W	-16.33	1.21	-15.12	-15.46	-2.71	-2.93	-8.26	-1.25
G:T	-18.07	2.05	-16.02	-16.42	-	-	-7.87	-
G:W	-9.43	0.59	-8.84	-8.57	7.18	7.85	-5.67	2.20

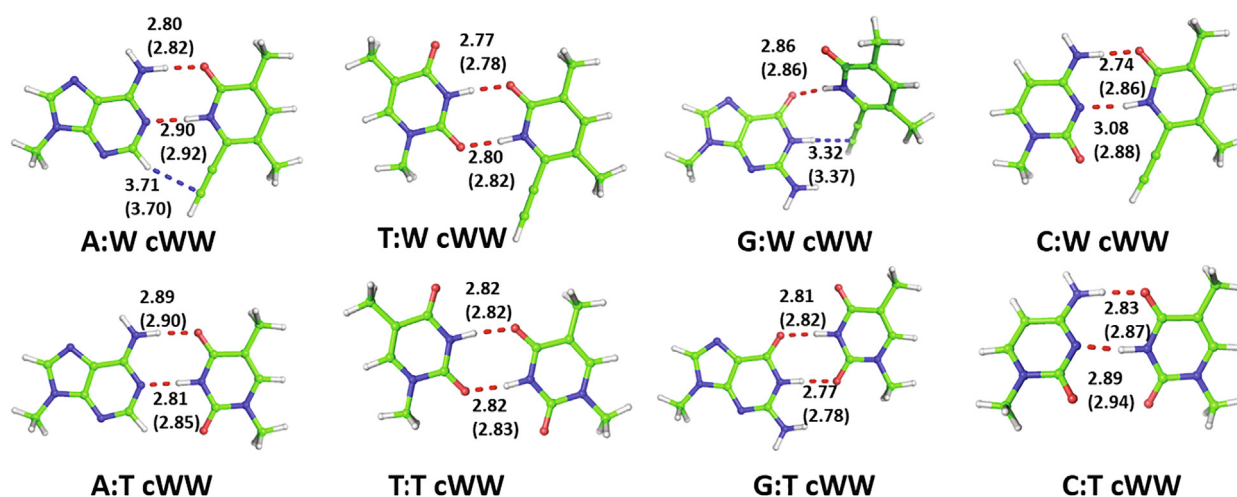


Fig. 2. Stick representation of the investigated base pairs. H-bond distances, in Å, are reported as red dashed lines for the optimized geometries in the gas phase (water). Interaction mediated by the ethynyl group of W are also indicated, by blue dashed lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

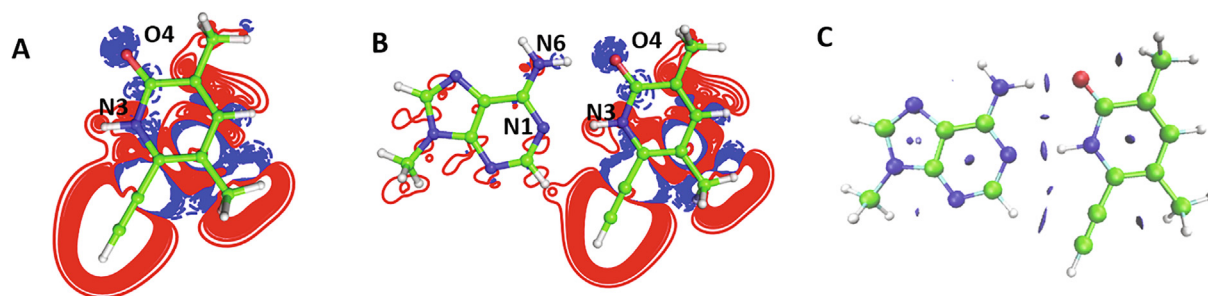


Fig. 3. Electron density difference, in the base plane, (A) between the W and T bases and (B) the A:W and A:T base pairs. Density difference curves are plotted between -0.02 and 0.02 a.u., with a spacing of 0.001 a.u. Blue (red) lines refer to negative (positive) density difference curves, i.e., to areas where the W base and AW base pair presents reduced (increased) electron density as compared to the system including T base and AT base pair. (C) NCI isosurface of the A:W base pair with representation of the reduced density gradient isosurface, $s = 0.5$ a.u. The surface is colored on a blue-to-red scale according to values of $\text{sign}(\lambda_2)\rho$, ranging from -0.5 to 0.5 a.u. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

E_{mod} of -2.71 and -1.25 kcal/mol in the gas phase and water, respectively.

3.1.3. G:W cWW base pair

The base pair formed by W with G results to be significantly propeller twisted both in the gas phase and in water. While maintaining C1'-C1' distances close to the canonical ones (10.84 Å in the gas phase and 10.99 Å in water), the optimized geometries deviate indeed significantly (by $52.8^\circ/51.8^\circ$, in the gas phase/water) from the planarity expected for a canonical cWW base pair. This is a

consequence of the steric repulsion between the G amino group and the W ethynyl group, which would crash in a planar cWW geometry.

In both the media, the H-bond between O6(G) and N3(W) is preserved, although being slightly elongated (by $0.04/0.05$ Å), and a weak attractive interaction seems to establish between the W ethynyl group and the N1(G) atom, as shown by the NCI analysis (see Fig. S1C).

Clearly, the G:W optimized geometries would not fit into a regular DNA duplex. Furthermore, they are weaker than the G:T base

pair, with E_{mod} values of 7.18 and 2.20 kcal/mol in the gas phase and in water, respectively.

3.1.4. C:W cWW base pair

In the gas phase optimization, C:W is stabilized by two strong H-bonds within a planar geometry, similarly to the natural C:T pair. The C1'–C1' distance is 8.94 Å/8.47 Å in the gas phase and in water, which compares with the distance of 8.52 Å/8.59 Å in C:T pair. In water optimization resulted in a optimized geometry substantially consistent with the gas phase one, with differences in the calculated H-bond distances within 0.20 Å and a C1'–C1' distance within 0.03 Å. As for the energy, the C:W pair results to be more stable than C:T, with E_{mod} of -4.64 kcal/mol and -2.16 kcal/mol in the gas phase and in water, respectively.

3.1.5. Reliability of calculated energetics of the H-bonded base pairs

In order to test the capability of the RIMP2 interaction energies to capture the modification of the heterocycle skeleton, along with its further functionalization of the non-natural nucleobase moiety (W base), we also calculated the gas phase DLPNO-CCSD(T)/CBS energies, considered to be the golden standard in quantum chemistry [30,31], for all the non-natural and unmodified base pairs under study. The values of calculated energies are reported in Table 1. Analogously to what recently reported [78], the comparison between the DLPNO-CCSD(T)/CBS energy values and those obtained by the RIMP2 approach shows that differences of the calculated energies are minor. They range between 0.1 kcal/mol for classical the A:T/G:C cWW base pairs and 0.7 kcal/mol for the non-natural base pairs under study. Further, differences in the corresponding E_{mod} values are small: 0.80 kcal/mol for A:W, 0.43 kcal/mol for C:W, 0.22 for T:W and 0.67 kcal/mol for G:W. This substantial agreement between the DLPNO-CCSD(T)/CBS and RIMP2/aug-cc-pVTZ calculated energy values further supports the usage of computationally less expensive RIMP2 energies in the context of H-bonded base pairs in nucleic acid structures. This is a relevant conclusion, as the DLPNO-CCSD(T) approach still is very expensive for very large systems, which are instead doable at the RIMP2 level. Further, the reported RIMP2 values allows to compare results in this work with a vast amount of literature in the field [24,28,78–85].

3.2. Dynamic behavior of a DNA duplex featuring a central A:W base pair

In order to investigate the effect of a central A:W pair on a DNA double helix, we simulated by classical MD the 12-mer DNA duplex previously proposed and experimentally characterized in [12] (see Fig. 1). We also simulated the same duplex with a central natural A:T pair, as a reference system, with the natural duplex being experimentally shown to feature a melting temperature of 4.4 °C lower than the corresponding modified one [12]. Each of the obtained MD trajectories was 500-ns long.

MD simulations showed that both the modified 1AW-DNA and unmodified 1AT-DNA duplexes maintain their overall topology during the simulations. The root mean square deviation (RMSD) profiles of all the atoms, excluding the capping base pairs, are fairly similar for the two systems with mean values of 2.68 ± 0.02 and 2.56 ± 0.04 Å, for the 1AW-DNA and the 1AT-DNA duplex, respectively (see Fig. 4A).

The distances between the heavy atoms of the central A:W and A:T base pairs potentially involved in inter-base H-bonds in the two systems have been monitored along the corresponding trajectories. Not surprisingly, the two strong H-bonds are retained in the 1AT-DNA duplex with average distances of 2.98 ± 0.01 and 2.94 ± 0.01 Å for N6(A)–O4(T) and N1(A)–N3(T), respectively, which are slightly larger than those obtained from the QM calculations.

The two expected H-bonds in 1AW-DNA, N6(A)–O4(W) and N1(A)–N3(W), are also maintained, with distances of 2.92 ± 0.01 and 2.99 ± 0.01 Å, respectively (Fig. S2), which are again only slightly larger than those obtained from the QM calculations. The C1'–C1' distance for the central base pair also remains stable under dynamic conditions, for both the 1AT-DNA and 1AW-DNA, with average values of 10.60 ± 0.02 Å and 10.96 ± 0.04 Å, respectively (Fig. S3). These values are also extremely close (deviations within 0.08 Å) to the ones obtained from the QM calculations. The interaction energies calculated by a molecular mechanics approach for the central A:T and A:W base pairs over the MD simulation time are -10.58 ± 0.04 kcal/mol and -12.28 ± 0.17 kcal/mol, respectively, being perfectly in line with the stability trend outlined by the QM calculations.

The glycosidic bond preference, i.e. the orientation of the nucleobase with respect to the furanose sugar, was also assessed for the central T7 and W7 nucleotides in the 1AT-DNA and 1AW-DNA duplexes. In either duplex, T/W samples conformations within the *anti* region (Fig. S4), the glycosidic angle being in the 181° – 295° / 151° – 266° ranges, with average values of 237° and 194° , respectively (Fig. S5). The glycosidic torsion of the complementary A nucleotide (A18) is also, not surprisingly, peculiar to an *anti* conformation, and ranges between 161° and 312° , irrespectively of the identity of the complementary nucleotide (T7 or W7), with average values of 246° / 250° for 1AW-DNA/1AT-DNA.

To assess the effect of the central A:W base pair on the global features of the DNA duplex, we calculated all the inter- and intra-base pair parameters describing the double helix structure. The values averaged over the final 400 ns of relative MD trajectories are reported in Table 2, where they are also compared to standard values mediated over four X-ray structures and from a long-timescale MD trajectory for a canonical 12-mer B-DNA duplex of sequence d(CGCGAATTCGCG)₂, the Drew–Dickerson dodecamer (DDD) [86,87].

From Table 2, it is clear that all the parameters of the modified duplex are close to those of a canonical B-DNA crystal structure, with maximum deviations in average angles and distances of 5.4° (for the buckle) and of 0.67 Å (for the slide). As for the comparison with the corresponding natural duplex, 1AT-DNA, deviations are, as expected, even smaller, with differences in angles within 2.4° and in distances within 0.2 Å.

The mean rise/twist parameters are 3.42 Å/ 33.3° for the 1AW-DNA system and 3.19 Å/ 32.9° for the 1AT-DNA system, respectively. These values are extremely close to those of a standard B-DNA, namely 3.35 Å for the rise and 35.2° for the twist; see Table 2 [87]. Their values are also consistent with those of the reference MD trajectory [86,87]. For example, the average intra-base propeller and inter-base rise values of $-10.3/-10.9^\circ$ and $3.42/3.19$ Å that we calculated both for the 1AW-DNA and the 1AT-DNA duplexes, are consistent with the values of approximately -10° and 3.3 Å from ref. [87], reported in Table 2. On a residue basis, it is clear from Fig. 5 that no considerable difference in the base-pair parameters for both 1AW-DNA and 1AT-DNA are observed near the modification site, except for the propeller and twist parameters, which deviate by ~ 10 – 15° for the base pair at the 5' side.

Furthermore, the spatial exploration of W relatively to the complementary A base in the 1AW-DNA is very similar to that of T in the 1AT-DNA duplex, indicating that the modification does not substantially change the dynamical behavior of the base pair in the context of a DNA duplex (see Fig. 6).

The rather small standard deviation for all the average structural parameters reported in Table 2 and for the RMSD of Fig. 4A clearly indicates that there is no drift in these parameters over time, confirming that the simulations can be considered equilibrated in the examined time window.

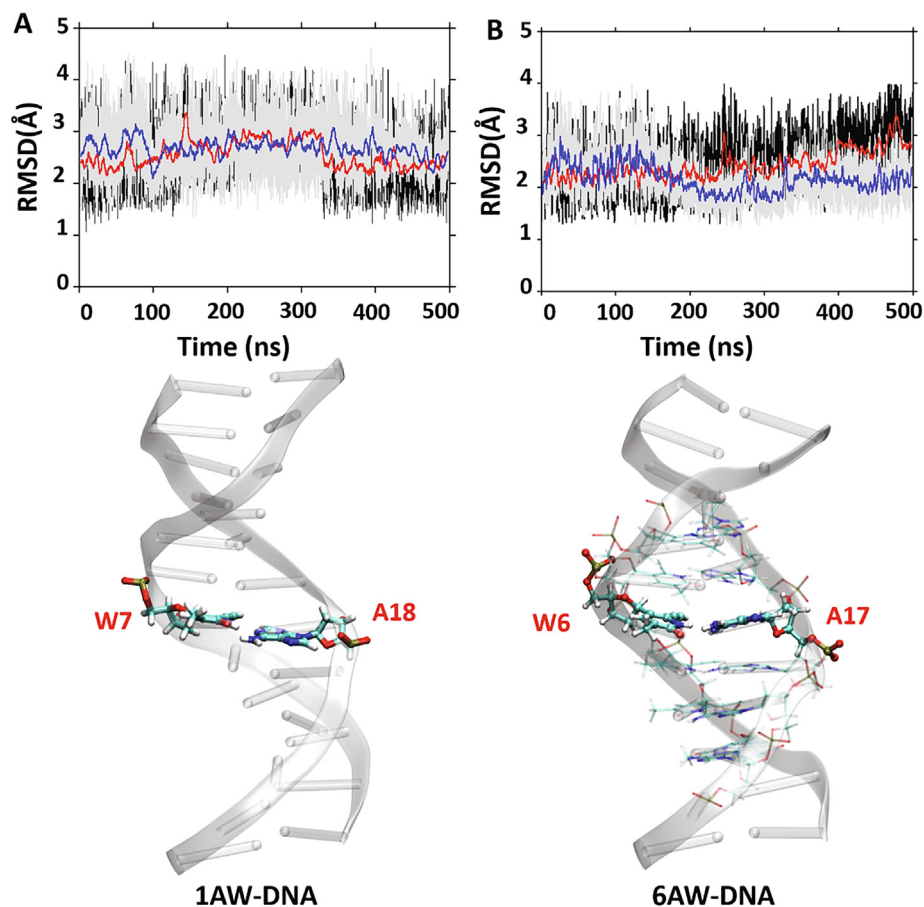


Fig. 4. A) *Top.* Root mean square deviation (RMSD) of the 1AW-DNA (grey and blue lines) and of the 1AT-DNA (black and red lines) systems, calculated on all the atoms but the capping base pairs. Blue and red lines correspond to rolling RMSD averaged over 500 MD frames. *Bottom.* Averaged 1AW-DNA structure shown in a ribbon representation; the central base pair, W7:A18, is shown in a stick representation. B) *Top.* RMSD of the 6AW-DNA (grey and blue lines) and of the 6AT-DNA (black and red lines) duplexes, calculated on all the atoms but the capping base pairs. Blue and red lines correspond to rolling RMSD averaged over 500 MD frames. *Bottom.* Averaged 6AW-DNA structure shown in a ribbon representation, with the central base pair, W6:A17, shown in a stick representation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

Helical parameters for the modified 1AW-DNA and 6AW-DNA duplexes and the corresponding unmodified 1AT-DNA and 6AT-DNA duplexes, averaged over the final 400 ns relative MD trajectories. In the last two columns, the base pair parameters from a long time scale dynamics and from X-ray structures of the Drew–Dickerson Dodecamer (DDD) [87] are reported for comparison.

	1AW-DNA	1AT-DNA	6AW-DNA	6AT-DNA	DDD_MD ^a	DDD_X-ray ^b
<i>Intra-Base Pair</i>						
Buckle (°)	-5.9 ± 0.3	-8.3 ± 0.3	-7.1 ± 0.2	-4.8 ± 0.3	0.0 ± 9.7	-0.5
Opening (°)	2.3 ± 0.3	2.0 ± 0.3	-0.7 ± 0.1	1.4 ± 0.5	1.3 ± 4.0	1.6
Propeller (°)	-10.3 ± 0.5	-10.9 ± 0.3	-11.6 ± 0.1	-8.7 ± 1.0	-9.2 ± 8.4	-14.4
Shear (Å)	0.04 ± 0.03	-0.01 ± 0.07	0.02 ± 0.05	-0.32 ± 0.10	0.0 ± 0.30	0.03
Stagger (Å)	0.02 ± 0.02	0.19 ± 0.05	0.84 ± 0.29	0.29 ± 0.24	0.10 ± 0.38	0.21
Stretch (Å)	-0.20 ± 0.05	-0.41 ± 0.07	-0.11 ± 0.06	-0.44 ± 0.11	0.02 ± 0.12	0.19
<i>Inter-Base Pair</i>						
Twist (°)	33.3 ± 0.2	32.9 ± 0.4	34.1 ± 0.1	33.2 ± 0.2	34.3 ± 5.5	35.2
Roll (°)	1.4 ± 0.3	2.2 ± 0.4	-0.4 ± 0.2	4.1 ± 0.8	1.5 ± 5.5	-0.7
Tilt (°)	1.3 ± 0.1	1.3 ± 0.4	1.9 ± 0.1	-0.1 ± 0.6	-0.1 ± 4.7	-0.4
Shift (Å)	0.22 ± 0.02	0.22 ± 0.04	0.21 ± 0.01	0.32 ± 0.05	-0.01 ± 0.81	-0.07
Slide (Å)	-0.53 ± 0.02	-0.43 ± 0.02	-0.03 ± 0.02	-0.50 ± 0.02	-0.24 ± 0.53	0.14
Rise (Å)	3.4 ± 0.10	3.2 ± 0.0	3.4 ± 0.0	3.3 ± 0.1	3.32 ± 0.29	3.35
<i>Base pair axis</i>						
Axbend (°)	18.8 ± 1.1	25.8 ± 0.8	17.3 ± 0.4	21.4 ± 2.0	-	-
Inclination (°)	2.7 ± 0.6	5.6 ± 0.5	-0.4 ± 0.2	4.9 ± 0.8	2.2 ± 5.7	-0.6
Tip (°)	-1.7 ± 0.3	0.1 ± 0.4	-2.9 ± 0.1	-3.8 ± 0.8	0.1 ± 6.9	-2.6
X-disp (Å)	-0.91 ± 0.07	-0.95 ± 0.02	0.00 ± 0.04	-1.19 ± 0.05	-0.58 ± 1.05	-0.15
Y-disp (Å)	-0.03 ± 0.03	-0.12 ± 0.03	-0.04 ± 0.01	-0.26 ± 0.05	0.00 ± 0.84	0.52

^a From a reference on long-timescale MD simulations [87].

^b Averaged over the X-ray structures with PDB IDs: 1BNA, 2BNA, 7BNA and 9BNA.

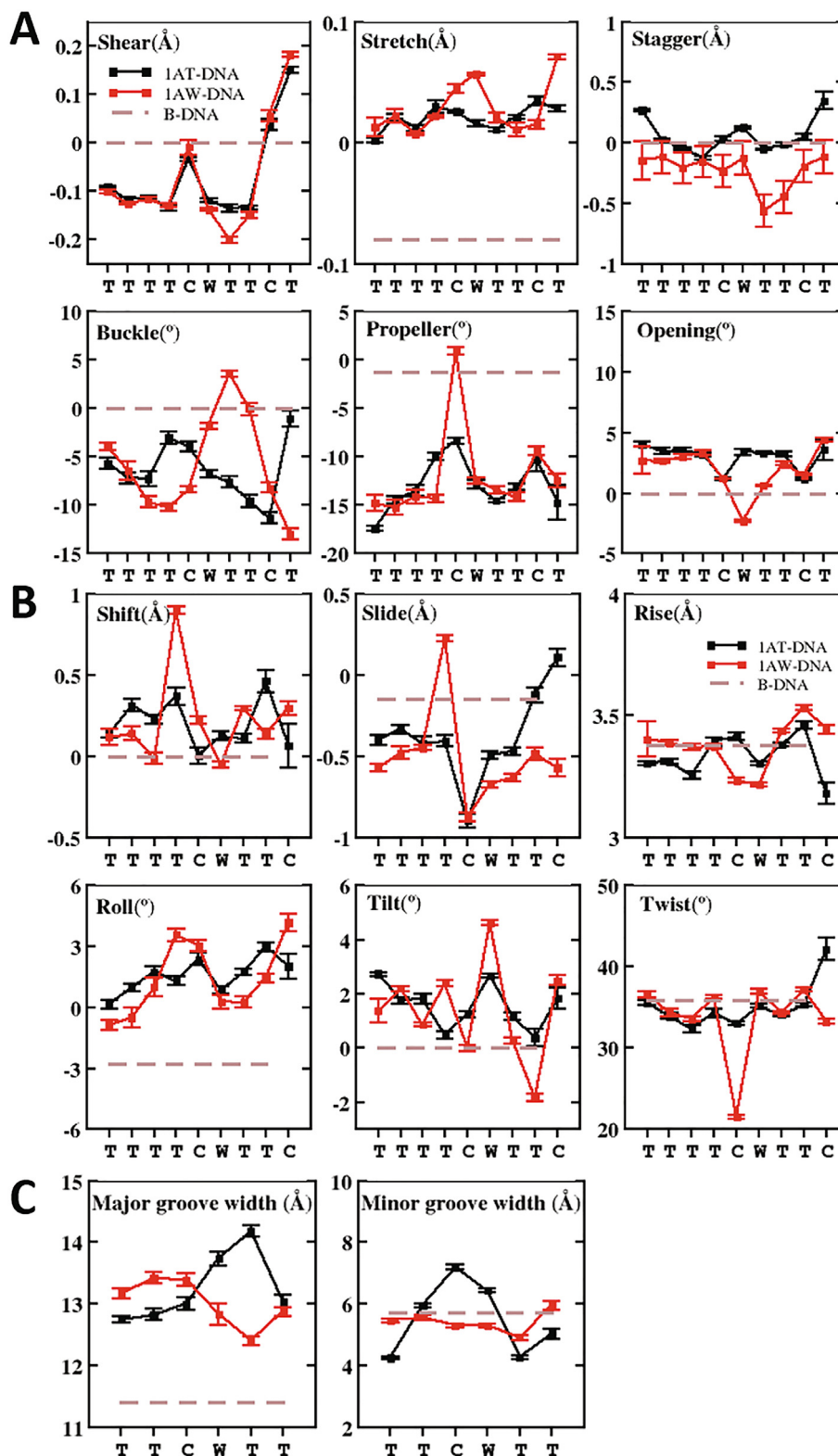


Fig. 5. Average values of the inter base-pair (A), and intra base-pair (B) helical parameters (C) groove parameters corresponding to the 1AW-DNA and the 1AT-DNA duplexes, calculated from the MD simulation trajectories. Translational parameters are in angstroms (Å) and rotational parameters in degrees (°). The error bars are the standard deviation.

Finally, it is well known that a B-DNA duplex possesses distinct grooves with a wide major groove and narrow minor groove. The width of the 1AW-DNA major and narrow minor grooves over the simulation time average to $13.03 \pm 0.08 \text{ \AA}$ and $5.25 \pm 0.04 \text{ \AA}$,

respectively (Fig. 5C), values which are very close to those of the unmodified 1AT-DNA (major and minor groove widths of $12.97 \pm 0.06 \text{ \AA}$ and $5.52 \pm 0.05 \text{ \AA}$, respectively). These values are also within 1.5/0.5 Å from those of a standard B-DNA duplex.

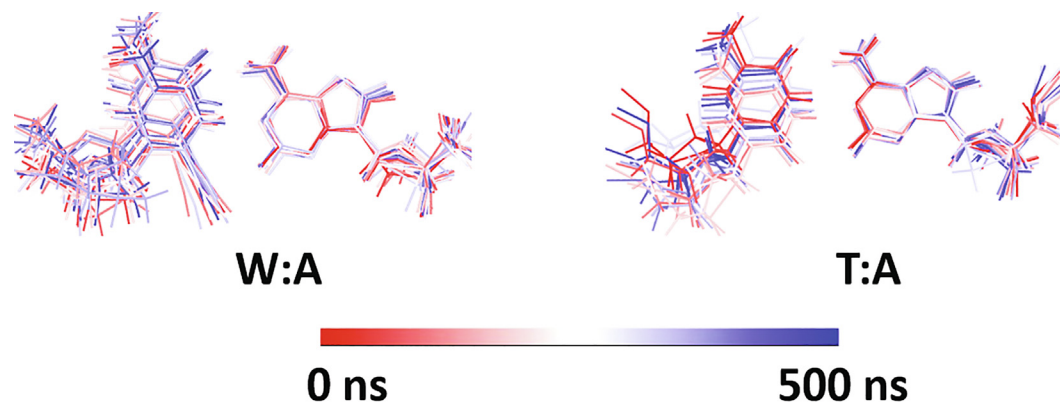


Fig. 6. Best superimposition of A in 20 representative snapshots of the WA and TA base pairs in the 1AW-DNA and 1AT-DNA duplexes, respectively, extracted from the MD simulations every 25 ns.

3.3. Dynamic behavior of a B-DNA duplex featuring multiple A:W base pairs.

To investigate the effect of multiple neighboring A:W base pairs on a DNA double strand, we simulated a model 11-mer DNA duplex by MD simulations, also experimentally characterized [12], featuring six consecutive A:W base pairs in its center (6AW-DNA, see Fig. 1). We simulated as well its unmodified counterpart, featuring 6 A:T base pairs in place of the modified A:W pairs (6AT-DNA) which has a lower melting temperature of 17.5 °C [12]. Each of the obtained MD trajectories was 500-ns long.

The RMSD profiles of 6AW-DNA and 6AT-DNA, calculated on all the atoms but the capping base pairs, are fairly similar, with mean values of 2.17 ± 0.03 and 2.43 ± 0.03 Å, respectively (see Fig. 4B). The H-bonding within all the six A:W base pairs were retained during the simulation time.

The C1'–C1' distance for the six central base pairs also remains stable under dynamic conditions, for both the 6AT-DNA and 6AW-DNA, with average values of 10.59 ± 0.01 Å and 10.95 ± 0.01 Å, respectively, from the MD trajectories. These values are also extremely close (deviations within 0.07 Å) to the ones obtained from the QM calculations.

Both in the 6AT-DNA and in the 6AW-DNA duplex, W and T sample conformations within the *anti* region. The average glycosidic torsion angles vary between 235° and 244° for the six central T bases in the 6AT-DNA system (Fig. S6), while they vary between 211° and 250° for the six central W bases in the 6AW-DNA system (Fig. S6). The glycosidic torsion values of the complementary A nucleotides are also, not surprisingly, peculiar to an *anti*-conformation, with average values in the 241–263° range for both the systems, irrespectively of the identity of the complementary nucleotide. The distribution of glycosidic torsion angles for the six central W/T nucleotides and their paired A nucleotides is reported in Figs. S7 and S8.

The inter- and intra-base pair parameters values averaged over the final 400 ns of relative MD trajectories are reported in Table 2 and compared to standard values for a canonical X-ray B-DNA structure and for a reference MD trajectory, as explained above.

From Table 2, it is clear that all the parameters of the modified 6AW-DNA duplex are close to those of a canonical B-DNA, with maximum deviations in average angles and distances of 6.6° (for the buckle) and of 0.63 Å (for the stagger). As for the comparison with the corresponding natural duplex, 6AT-DNA, deviations are, as expected, even smaller, with differences in angles within 4° (for roll) and in distances within 0.5 Å (for stagger), see Table 2.

The mean rise/twist parameters are 3.4 Å/34.1° for the 6AW-DNA system and 3.3 Å/33.2° for the 6AT-DNA system, respectively. These values are extremely close to those of a standard B-form DNA, namely 3.32 Å for the rise and 34.3° for the twist [87]. In addition, average structural parameters in Table 2 are also consistent with those reported for the reference B-DNA MD trajectory [86]. For example, the average intra-base propeller and inter-base rise values of $-11.6/-8.7^\circ$ and 3.30/3.32 Å that we calculated for both the 6AW- and 6AT-DNA duplexes are consistent with the values of approximately -10° and 3.3 Å reported in ref. [86] (see also Table 2). On a residue basis, it is clear from Fig. 7 that no considerable difference between corresponding base-pair parameters in 6AW-DNA and 6AT-DNA is observed.

As compared to 1AW-DNA, 6AW-DNA shows moderately higher stretch and shear values (Fig. S9). Consistently higher stagger values (within 1.0 Å) are also observed for all the base pairs in 6AW-DNA, as compared to 1AW-DNA and the standard B-DNA duplex.

Finally, the major and minor groove widths over the simulation time average to 12.96 ± 0.11 Å and 4.44 ± 0.09 Å, respectively, which values very close (within 0.55 Å) to those of the corresponding unmodified 6AT DNA (with major and minor groove widths of 12.48 ± 0.13 Å and 4.99 ± 0.06 Å, respectively, Fig. 7C). These values are also within 1.5 Å from those of a standard B-DNA duplex.

3.4. Stacking energies of the central bases in the 1AW/6AW-DNA and 1AT/6AT-DNA duplex

Geometries corresponding to five snapshots extracted every 100 ns from the MD simulations were also used to investigate, by a QM approach, the impact of the non-natural W base on the stacking with adjacent nucleobases in DNA duplexes.

For 1AW-DNA, we calculated the stacking energies of the single bases in the central pair with the stacked bases in the 3' and in the 5' directions, for a total of 4 stacked bases, 2 for 1AW-DNA, 2 for 1AT-DNA (see Fig. 8). Consistently with the experimental evidence, our QM calculations indicate a stabilization of 2.53 kcal/mol when the 3' W7 base is stacked on the 5' C6 base, as compared to the corresponding natural 5'-C6//T7-3' stacking interaction in the 1AT-DNA duplex, and similar stabilization for the W base stacked on the 3' T8 base in 1AW-DNA, as compared to the corresponding 5'-T7//T8-3' interaction in 1AT-DNA (Fig. 8A, B).

For 6AW-DNA, stacking energies for a total of 6 stacked base pairs, 3 for 6AW-DNA and 3 for 6AT-DNA, were calculated, at the central positions 5//6, and at the positions immediately down and upstream to the modifications, 3//4 and 10//11. The stacking energy of the two consecutive W bases giving the 5'-W5//W6-3'

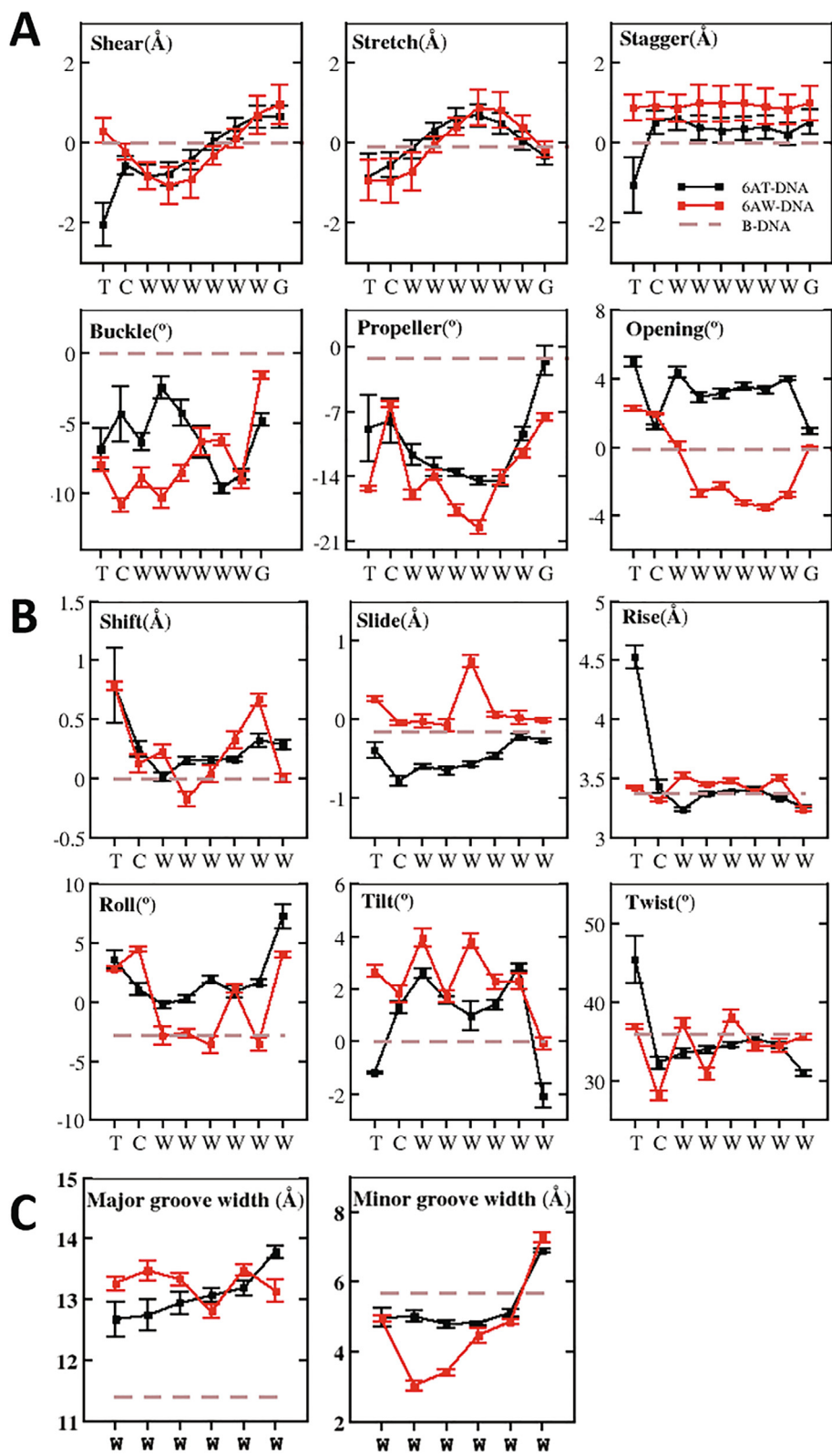


Fig. 7. Average values of the inter-base-pair (A), and intra-base-pair (B) helical parameters (C) groove parameters in 6AT-DNA and 6AW-DNA duplexes calculated from the MD trajectories. Translational parameters are in angstroms (Å) and rotational parameters in degrees (°). The error bars are the standard deviation.

interaction is more stable by ~1 kcal/mol as compared to the 5'-T5//T6-3' interaction in 6AT-DNA (Fig. 8C, D). Further, the 5'-C3//W4-3' 3'interaction is more stable by ~0.5 kcal/mol as compared to the 5'-C3//T4-3' interaction in 6AT-DNA (Fig. 8E, F).

Finally, a similar stability is observed for the 5'-W9//G10-3' stacking interaction and the corresponding 5'-T9//G10-3' interaction in 6AT-DNA (with a difference in energy of only ~0.04 kcal/mol; Fig. 8G, H).

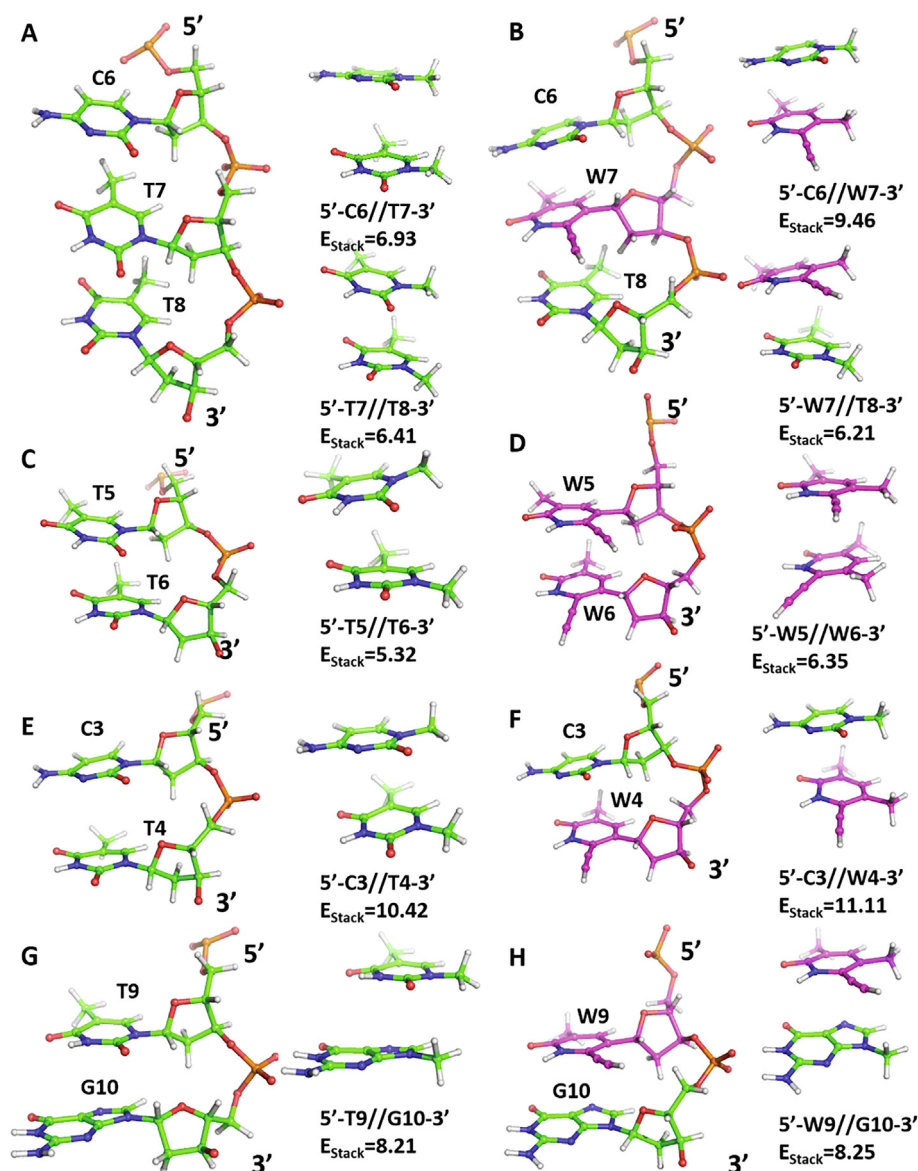


Fig. 8. Side view of the three central bases in the 1AT-DNA (A) and 1AW-DNA (B) duplexes. The side views of the stacked base//base pairs in the 1AT-DNA (A) and 1AW-DNA (B) systems. Side view of the two central bases in the 6AT-DNA (C) and 6AW-DNA (D) duplexes. The side views of the stacked base//base pairs in the 6AT-DNA (C, E, G) and 6AW-DNA (D, F, H) systems with the corresponding gas phase interaction energies, E_{Stack} , in kcal/mol. Nucleotides are numbered according to the scheme proposed in [12], see Fig. 1.

The above findings thus point overall to a stabilizing contribution from the non-natural W base to the stacking with neighboring bases, for both the duplexes.

4. Conclusions

The T analogue abbreviated as W, bearing an ethynyl group at the C2 position in place of the T carbonyl oxygen, proposes itself as an ideal fifth base for biotechnological applications where a comparable stability of the base pairs, independently from the DNA/RNA sequence, is wanted. It was experimentally shown to be incorporated in nucleotide sequences by automated synthesis in high yield, to exhibit high pairing fidelity and to introduce thermal stabilization over the corresponding A:T containing duplexes.

Herein, by an integrated QM and MD approach, we dissected and explained the effect of the W modification on the base pairing with A and the other natural bases, and on stacking interactions

with neighboring bases; furthermore, we explored theoretically the compatibility of single and multiple W insertions with the dynamical features of B-DNA duplexes.

First of all, our QM calculations revealed that the optimized W:G pair results in a geometry distorted in its planarity and incompatible with a regular DNA duplex, thus explaining the severe depression in the UV-melting point, by 20.5 °C, for duplexes carrying even a single G:W mismatch [12] and the high base pairing fidelity of the W base.

The A:W base pair, in its classical *cis* Watson-Crick geometry, was shown instead to preserve planarity and to exhibit a CH- π interaction between the C2-H on A and the ethynyl group on W, in addition to the two A:T hydrogen bonds, which are preserved. This results in a stronger modified A:W base pair, as compared to the unmodified A:T one.

MD simulations on two duplexes, 1AW-DNA and 6AW-DNA, carrying one and six consecutive W bases, respectively, paired to complementary A bases, showed that they preserve the

anti-conformation of their glycosidic bonds and, overall, the structural parameters of a B-DNA duplex. Only minor changes are observed as compared to the corresponding unmodified duplexes and also to a standard B-DNA (the Drew-Dickerson dodecamer) used as a golden standard to compare both static and dynamical DNA helix parameters [86,87]. Within these minor changes, there is a propeller-twist by $\sim 10\text{--}15^\circ$ for the base pair at the 5' side of the modified W in 1AW-DNA.

Both the 1AW-DNA and 6AW-DNA exhibit distinctly different groove widths, as expected for a B-DNA. The difference in the major and minor groove sizes is even enhanced in the modified duplexes, with average changes in widths within 1.5 Å both from the corresponding unmodified duplexes and from the 'standard' Drew-Dickerson dodecamer. The maximum local deviation is observed for the minor groove, which is shrunk by ~ 2 Å in the middle of the 6AW-DNA duplex, as compared to the unmodified 6AT-DNA. These moderate deviations are still compatible with a B-DNA duplex.

Snapshots of adjacent base pairs, extracted from the MD simulations of the duplexes, allowed us to also investigate, by a QM approach, the energy of the relative stacking interactions, which resulted consistently stronger for W as compared to T. Therefore, the enhanced stability of duplexes including single or multiple W bases in place of A can be explained with the enhanced stability of the modified A:W base pairs, along with the stronger stacking interaction of W with neighboring bases. Importantly, this gained stability is achieved while maintaining the canonical geometrical features of a B-DNA, even under dynamical conditions, thus presumably preserving the interaction with physiological partners in the cell; this holding true also for the insertion of several consecutive W bases.

In conclusion, our study confirms W as an extremely promising fifth base for biotechnological applications, while explaining in detail the reasons for its stabilizing effect.

CRediT authorship contribution statement

Mohit Chawla: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing - original draft, Writing - review & editing, Validation, Visualization. **Suresh Gorle:** Data curation, Formal analysis, Methodology, Validation, Visualization. **Abdul Rajjak Shaikh:** Data curation, Methodology. **Romina Oliva:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing - original draft, Writing - review & editing. **Luigi Cavallo:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

L.C. and M. C. acknowledges King Abdullah University of Science and Technology (KAUST) for support and the KAUST Supercomputing Laboratory for providing computational resources of the supercomputer Shaheen II. R.O. thanks MIUR-FFABR (Fondo per il Finanziamento Attività Base di Ricerca) for funding.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2021.02.006>.

References

- [1] Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, Olchowiak A, et al. MODOMICS: a database of RNA modification pathways—2013 update. *Nucleic Acids Res* 2013;41:D262–267.
- [2] Egger G, Liang GN, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 2004;429:457–63.
- [3] Malyshev DA, Dhama K, Lavergne T, Chen T, Dai N, Foster JM, et al. A semi-synthetic organism with an expanded genetic alphabet. *Nature* 2014;509:385–8.
- [4] Phelps K, Morris A, Beal PA. Novel modifications in RNA. *ACS Chem Biol* 2012;7:100–9.
- [5] Xu W, Chan KM, Kool ET. Fluorescent nucleobases as tools for studying DNA and RNA. *Nat Chem* 2017;9:1043–55.
- [6] Li L, Degardin M, Lavergne T, Malyshev DA, Dhama K, Ordoukhanian P, et al. Natural-like replication of an unnatural base pair for the expansion of the genetic alphabet and biotechnology applications. *J Am Chem Soc* 2014;136:826–9.
- [7] Hoshika S, Leal NA, Kim M-J, Kim M-S, Karalkar NB, Kim H-J, et al. Hachimoji DNA and RNA: a genetic system with eight building blocks. *Science* 2019;363(6429):884–7.
- [8] Georgiadis MM, Singh I, Kellett WF, Hoshika S, Benner SA, Richards NGJ. Structural basis for a six nucleotide genetic alphabet. *J Am Chem Soc* 2015;137:6947–55.
- [9] Dhama K, Malyshev DA, Ordoukhanian P, Kubelka T, Hocek M, Romesberg FE. Systematic exploration of a class of hydrophobic unnatural base pairs yields multiple new candidates for the expansion of the genetic alphabet. *Nucleic Acids Res* 2014;42:10235–44.
- [10] Hirao I, Kimoto M, Mitsui T, Fujiwara T, Kawai R, Sato A, et al. An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA. *Nat Methods* 2006;3:729–35.
- [11] Eyberg J, Richert C. Glen report 31.21: ethynylpyridone C-nucleoside phosphoramidite (dW): a high affinity replacement for thymidine. glenresearch.com, 2020.
- [12] Walter TJ, Richert C. A strongly pairing fifth base: oligonucleotides with a C-nucleoside replacing thymidine. *Nucleic Acids Res* 2018;46:8069–78.
- [13] Sintim HO, Kool ET. Enhanced base pairing and replication efficiency of thiothymidines, expanded-size variants of thymidine. *J Am Chem Soc* 2006;128:396–7.
- [14] Liu HB, Gao JM, Lynch SR, Saito YD, Maynard L, Kool ET. A four-base paired genetic helix with expanded size. *Science* 2003;302:868–71.
- [15] Lin KY, Matteucci MD. A cytosine analogue capable of clamp-like binding to a guanine in helical nucleic acids. *J Am Chem Soc* 1998;120:8531–2.
- [16] Chaudhuri NC, Kool ET. Very high-affinity DNA recognition by bicyclic and cross-linked oligonucleotides. *J Am Chem Soc* 1995;117:10434–42.
- [17] Wagner RW, Matteucci MD, Lewis JG, Gutierrez AJ, Moulds C, Froehler BC. Antisense gene inhibition by oligonucleotides containing C-5 propyne pyrimidines. *Science* 1993;260:1510–3.
- [18] Minuth M, Richert C. A nucleobase analogue that pairs strongly with adenine. *Angew Chem Int Ed* 2013;52:10874–7.
- [19] Halder A, Datta A, Bhattacharyya D, Mitra A. Why does substitution of thymine by 6-ethynylpyridone increase the thermostability of DNA double helices?. *J Phys Chem B* 2014;118:6586–96.
- [20] Gibson DJ, van Mourik T. Stacking with the unnatural DNA base 6-ethynylpyridone. *Chem Phys Lett* 2017;668:7–13.
- [21] Ivarie R. Thymine methyls and DNA-protein interactions. *Nucleic Acids Res* 1987;15:9975–83.
- [22] Chawla M, Abdel-Azeim S, Oliva R, Cavallo L. Higher order structural effects stabilizing the reverse Watson-Crick guanine-cytosine base pair in functional RNAs. *Nucleic Acids Res* 2014;42:714–26.
- [23] Chawla M, Credendino R, Oliva R, Cavallo L. Structural and energetic impact of non-natural 7-deaza-8-azaadenine and its 7-substituted derivatives on H-bonding potential with uracil in RNA molecules. *J Phys Chem B* 2015;119:12982–9.
- [24] Chawla M, Oliva R, Bujnicki JM, Cavallo L. An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies. *Nucleic Acids Res* 2015;43:6714–29.
- [25] Chawla M, Poater A, Besalu-Sala P, Kalra K, Oliva R, Cavallo L. Theoretical characterization of sulfur-to-selenium substitution in an emissive RNA alphabet: impact on H-bonding potential and photophysical properties. *Phys Chem Chem Phys* 2018;20:7676–85.
- [26] Chawla M, Poater A, Oliva R, Cavallo L. Structural and energetic characterization of the emissive RNA alphabet based on the isothiazolo[4,3-d]pyrimidine heterocycle core. *Phys Chem Chem Phys* 2016;18:18045–53.
- [27] Chawla M, Sharma P, Halder S, Bhattacharyya D, Mitra A. Protonation of base pairs in RNA: context analysis and quantum chemical investigations of their geometries and stabilities. *J Phys Chem B* 2011;115:1469–84.
- [28] Sponer J, Jurecka P, Hobza P. Accurate interaction energies of hydrogen-bonded nucleic acid base pairs. *J Am Chem Soc* 2004;126:10142–51.

- [29] Sponer J, Bussi G, Krepl M, Banas P, Bottaro S, Cunha RA, et al. RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chem Rev* 2018;118:4177–338.
- [30] Kruse H, Banas P, Sponer J. Investigations of stacked DNA base-pair steps: highly-accurate stacking interaction energies, energy decomposition and many-body stacking effects. *J Chem Theor Comput* 2018;15:95–115.
- [31] Kruse H, Mladek A, Gkionis K, Hansen A, Grimme S, Sponer J. Quantum chemical benchmark study on 46 RNA backbone families using a dinucleotide unit. *J Chem Theor Comput* 2015;11:4972–91.
- [32] Chawla M, Autiero I, Oliva R, Cavallo L. Energetics and dynamics of the non-natural fluorescent 4AP:DAP base pair. *Phys Chem Chem Phys* 2018;20:3699–709.
- [33] Chawla M, Chermak E, Zhang QY, Bujnicki JM, Oliva R, Cavallo L. Occurrence and stability of lone pair- π stacking interactions between ribose and nucleobases in functional RNAs. *Nucleic Acids Res* 2017;45:11019–32.
- [34] Chawla M, Credendino R, Poater A, Oliva R, Cavallo L. Structural stability, acidity, and halide selectivity of the fluoride riboswitch recognition site. *J Am Chem Soc* 2015;137:299–306.
- [35] Sponer J, Leszczynski J, Hobza P. Nature of nucleic acid-base stacking: nonempirical ab initio and empirical potential characterization of 10 stacked base dimers. Comparison of stacked and H-bonded base pairs. *J Phys Chem* 1996;100:5590–6.
- [36] Sponer JE, Leszczynski J, Sychrovsky V, Sponer J. Sugar edge/sugar edge base pairs in RNA: stabilities and structures from quantum chemical calculations. *J Phys Chem B* 2005;109:18680–9.
- [37] Kalra K, Gorle S, Cavallo L, Oliva R, Chawla M. Occurrence and stability of lone pair- π and OH- π interactions between water and nucleobases in functional RNAs. *Nucleic Acids Res* 2020;48:5825–38.
- [38] Flamme M, Rothlisberger P, Levi-Acobas F, Chawla M, Oliva R, Cavallo L, et al. Enzymatic formation of an artificial base pair using a modified purine nucleoside triphosphate. *ACS Chem Biol* 2020;15:2872–84.
- [39] Becke AD. density-functional thermochemistry. 3. The Role of exact exchange. *J Chem Phys* 1993;98:5648–52.
- [40] Becke AD. Density-functional thermochemistry. Abs Pap Am Chem Soc 1996;212:112-COMP.
- [41] Lee CT, Yang WT, Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron-density. *Phys Rev B* 1988;37:785–9.
- [42] Dunning TH. Gaussian-basis sets for use in correlated molecular calculations. 1. The atoms boron through neon and hydrogen. *J Chem Phys* 1989;90:1007–23.
- [43] Klamt A, Schüürmann G. Cosmo – a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J Chem Soc* 1993;2:799–805.
- [44] Grimme S, Antony J, Ehrlich S, Krieg H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J Chem Phys* 2010;132:154104.
- [45] Moller C, Plesset MS. Note on an approximation treatment for many-electron systems. *Phys Rev* 1934;46:0618–22.
- [46] Weigend F, Haser M. RI-MP2: first derivatives and global consistency. *Theor Chem Acc* 1997;97:331–40.
- [47] Boys SF, Bernardi F. Calculation of small molecular interactions by differences of separate total energies – some procedures with reduced errors. *Mol Phys* 1970;19:553.
- [48] Sponer JE, Reblova K, Mokdad A, Sychrovsky V, Leszczynski J, Sponer J. Leading RNA tertiary interactions: structures, energies, and water insertion of A-minor and P-interactions. A quantum chemical view. *J Phys Chem B* 2007;111:9153–64.
- [49] Sponer JE, Spackova N, Kulhanek P, Leszczynski J, Sponer J. Non-Watson-Crick base pairing in RNA. Quantum chemical analysis of the cis Watson-Crick/sugar edge base pair family. *J Phys Chem A* 2005;109:2292–301.
- [50] Sponer JE, Spackova N, Leszczynski J, Sponer J. Principles of RNA base pairing: structures and energies of the trans Watson-Crick/sugar edge base pairs. *J Phys Chem B* 2005;109:11399–410.
- [51] Sharma P, Sharma S, Chawla M, Mitra A. Modeling the noncovalent interactions at the metabolite binding site in purine riboswitches. *J Mol Model* 2009;15:633–49.
- [52] Oliva R, Cavallo L, Tramontano A. Accurate energies of hydrogen bonded nucleic acid base pairs and triplets in tRNA tertiary interactions. *Nucleic Acids Res* 2006;34:865–79.
- [53] Zirbel CL, Sponer JE, Sponer J, Stombaugh J, Leontis NB. Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res* 2009;37:4898–918.
- [54] Riplinger C, Neese F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J Chem Phys* 2013;138:034106.
- [55] Riplinger C, Sandhoefer B, Hansen A, Neese F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J Chem Phys* 2013;139:134101.
- [56] Riplinger C, Pinski P, Becker U, Valeev EF, Neese F. Sparse mads – a systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J Chem Phys* 2016;144:024109.
- [57] Neese F. Software update: the ORCA program system, version 4.0. *WIREs Comput Mol Sci* 2018;8.
- [58] Liakos DG, Sparta M, Kesharwani MK, Martin JML, Neese F. Exploring the accuracy limits of local pair natural orbital coupled-cluster theory. *J Chem Theory Comput* 2015;11:1525–39.
- [59] Weigend F, Kohn A, Hattig C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *J Chem Phys* 2002;116:3175–83.
- [60] Halkier A, Helgaker T, Jørgensen P, Klopper W, Koch H, Olsen J, et al. Basis-set convergence in correlated calculations on Ne, N-2, and H₂O. *Chem Phys Lett* 1998;286:243–52.
- [61] Helgaker T, Klopper W, Koch H, Noga J. Basis-set convergence of correlated calculations on water. *J Chem Phys* 1997;106:9639–46.
- [62] Halkier A, Helgaker T, Jørgensen P, Klopper W, Olsen J. Basis-set convergence of the energy in molecular Hartree-Fock calculations. *Chem Phys Lett* 1999;302:437–46.
- [63] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph Model* 1996;14:33–8.
- [64] Ivani I, Dans PD, Noy A, Perez A, Faustino I, Hospital A, et al. Parmbsc1: a refined force field for DNA simulations. *Nat Methods* 2016;13:55–8.
- [65] Vanquelf E, Simon S, Marquant G, Garcia E, Klimerak G, Delepine JC, et al. R.E. D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res* 2011;39:W511–517.
- [66] Dupradeau FY, Pigache A, Zaffran T, Savineau C, Lelong R, Grivel N, et al. The R. ED. tools: advances in RESP and ESP charge derivation and force field library building. *Phys Chem Chem Phys* 2010;12:7821–39.
- [67] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J Am Chem Soc* 1995;117:5179–97.
- [68] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. *J Comput Chem* 2005;26:1781–802.
- [69] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;79:926–35.
- [70] Berendsen HJC, Grigera JR, Straatsma TP. The missing term in effective pair potentials. *J Phys Chem* 1987;91:6269–71.
- [71] Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical-integration of Cartesian equations of motion of a system with constraints – molecular-dynamics of N-alkanes. *J Comput Phys* 1977;23:327–41.
- [72] The PyMOL molecular graphics system Schrödinger, LLC; 2016.
- [73] Racine J. gnuplot 4.0: a portable interactive plotting utility. *J Appl Econ* 2006;21:133–41.
- [74] Blanchet C, Pasi M, Zakrzewska K, Lavery R. CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res* 2011;39:W68–73.
- [75] Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA* 2001;7:499–512.
- [76] Contreras-Garcia J, Johnson ER, Keinan S, Chaudret R, Piquemal JP, Beratan DN, et al. NCIPLOT: a program for plotting non-covalent interaction regions. *J Chem Theor Comput* 2011;7:625–32.
- [77] Johnson ER, Keinan S, Mori-Sanchez P, Contreras-Garcia J, Cohen AJ, Yang W. Revealing noncovalent interactions. *J Am Chem Soc* 2010;132:6498–506.
- [78] Chawla M, Minenkov Y, Vu KB, Oliva R, Cavallo L. Structural and energetic impact of non-natural 7-deaza-8-azaguanine, 7-deaza-8-azaisoguanine, and their 7-substituted derivatives on hydrogen-bond pairing with cytosine and isocytosine. *ChemBioChem* 2019;20:2262–70.
- [79] Jurecka P, Hobza P. True stabilization energies for the optimal planar hydrogen-bonded and stacked structures of guanine...cytosine, adenine...thymine, and their 9- and 1-methyl derivatives: complete basis set calculations at the MP2 and CCSD(T) levels and comparison with experiment. *J Am Chem Soc* 2003;125:15608–13.
- [80] Riley KE, Hobza P. Assessment of the MP2 method, along with several basis sets, for the computation of interaction energies of biologically relevant hydrogen bonded and dispersion bound complexes. *J Phys Chem A* 2007;111:8257–63.
- [81] Khakshoor O, Wheeler SE, Houk KN, Kool ET. Measurement and theory of hydrogen bonding contribution to isosteric DNA base pairs. *J Am Chem Soc* 2012;134:3154–63.
- [82] Rejnek J, Hobza P. Hydrogen-bonded nucleic acid base pairs containing unusual base tautomers: complete basis set calculations at the MP2 and CCSD (T) levels. *J Phys Chem B* 2007;111:641–5.
- [83] Dabkowska I, Gonzalez HV, Jurecka P, Hobza P. Stabilization energies of the hydrogen-bonded and stacked structures of nucleic acid base pairs in the crystal geometries of CG, AT, and AC DNA steps and in the NMR geometry of the 5'-d(CCGAAGC)-3' hairpin: complete basis set calculations at the MP2 and CCSD(T) levels. *J Phys Chem A* 2005;109:1131–6.
- [84] Dabkowska I, Jurecka P, Hobza P. On geometries of stacked and H-bonded nucleic acid base pairs determined at various DFT, MP2, and CCSD(T) levels up to the CCSD(T)/complete basis set limit level. *J Chem Phys* 2005;122:204322.
- [85] Molt RW, Georgiadis MM, Richards NGJ. Consecutive non-natural PZ nucleobase pairs in DNA impact helical structure as seen in 50 mu s molecular dynamics simulations. *Nucleic Acids Res* 2017;45:3643–53.
- [86] Galindo-Murillo R, Robertson JC, Zgarbova M, Sponer J, Otyepka M, Jurecka P, et al. Assessing the current state of amber force field modifications for DNA. *J Chem. Theo. Comput.* 2016;12:4114–27.
- [87] Dans PD, Danilane L, Ivani I, Drasata T, Lankas F, Hospital A, et al. Long-timescale dynamics of the Drew-Dickerson dodecamer. *Nucleic Acids Res* 2016;44:4052–66.