Semi-empirical and Linear-Scaling DFT Methods to Characterize duplex DNA and G-quadruplexes in Presence of Interacting Small Molecules

Iker Ortiz de Luzuriaga,[†,‡] Sawssen Elleuchi,[¶] Emilio Artacho,[§,∥,†,⊥] Khaled Jarraya,[¶] Xabier Lopez,[*,⊥,‡] and Adrià Gil[*,†,#,@,△]

†CICnanoGUNE BRTA, Tolosa Hiribidea 76, E-20018, Donostia - San Sebastian.

‡Polimero eta Material Aurreratuak: Fisika, Kimika eta Teknologia, Kimika Fakultatea, Euskal Herriko Uniberstitatea, UPV/EHU, 20080 Donostia, Euskadi, Spain.

¶Laboratoire de Chimie Inorganique, LR17ES07, Université de Sfax, Faculté de Sciences de Sfax, 3000 Sfax, Tunisia.

§Theory of Condensed Matter, Cavendish Laboratory, University of Cambridge, J. J. Thomson Ave., Cambridge CB3 0HE, United Kingdom

∥Ikerbasque, Basque Foundation for Science, 48011 Bilbao, Spain

⊥Donostia International Physics Center, 20018 Donostia, Spain.

#ARAID Foundation, Zaragoza, Spain

@Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH) CSIC- Universidad de Zaragoza, C/ Pedro Cerbuna 12, 50009, Zaragoza, Spain

△BioISI – Biosystems and Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016, Lisboa, Portugal

E-mail: xabier.lopez@ehu.es;

adria.gil.mestres@csic.es;adriagilmestres@unizar.es;agmestres@fr.ul.pt

Abstract

The computational study of DNA and its interaction with ligands is a highly relevant area of research, with significant consequences for developing new therapeutic strategies. However, the computational description of such large and complex systems requires considering interactions of different types simultaneously in a balanced way, such as non-covalent weak interactions (namely hydrogen bonds and stacking), metal-ligand interactions, polarization and charge transfer effects. All these considerations imply a real challenge for computational chemistry. The possibility of studying large biological systems using quantum methods for the entire system requires significant computational resources, with improvements in parallelization and optimization of theoretical strategies. Computational methods, such as Linear-Scaling Density Functional Theory and DLPNO-CCSD(T), may allow performing *ab initio* QM calculations, including explicitly the electronic structure for large biological systems, at a reasonable computing time. In this work, we study the interaction of small molecules and cations with DNA (both duplex-DNA and G-quadruplexes), comparing different computational methods: a linear-scaling DFT (LS-DFT) at LMKLL/DZDP level of theory, semi-empirical methods (PM6-DH2 and PM7), mixed QM/MM, and DLPNO-CCSD(T). Our goal is to demonstrate the adequacy of LS-DFT to treat the different types of interactions present in DNA-dependent systems. We show that LMKLL/DZDP using SIESTA can yield very accurate geometries and energetics in all the different systems considered in this work: duplex DNA (dDNA), phenanthroline intercalating dDNA, G-quadruplexes, and Metal-G-tetrads considering alkaline metals of different sizes. As far as we know, this is the first time that full G-quadruplex geometry optimizations have been carried out using a DFT method thanks to its linear-scaling capabilities. Moreover, we show that LS-DFT provides high-quality structures, and some semi-empirical Hamiltonian can also yield suitable geometries. However, DLPNO-CCSD(T) and LS-DFT are the only methods that accurately describe interaction energies for all the systems considered in our study.

Introduction

The interaction of ligands with DNA is a vital research subject with critical therapeutic consequences. Small molecules have shown enough antitumoral activity[1] binding efficiently to duplex DNA (dDNA) or G-quadruplexes (GQ). For instance, cis-platin[2] is an effective drug which effectively binds to dDNA. However, the toxicity,[3,4] resistance and nonspecific interactions of available drugs make desirable the quest for new molecules that target DNA, in its canonical dDNA form or in other secondary DNA structures, such as G-quadruplexes. In this sense, theoretical methods that help elucidate the nature of the interactions between ligands and DNA are of paramount importance. Studies on the interaction of ligands with dDNA structures are available in the literature.[5–8] However, studies with other DNA structures, such as GQ, are scarcer.[9,10] GQ DNA structures are formed by the stacking of G-tetrads, each one composed by the planar arrangement of four guanine bases.[11] GQ have raised considerable interest during the past years for the development of therapies against cancer. These non-canonical structures of DNA may be found in telomeres and/or oncogene promoters, and it has been observed that the stabilization of such GQ may disturb tumor cell growth.[12,13]

There are different theoretical methods available for the study of DNA and the state-of-the-art for the GQ modeling has been reviewed very recently in our team.[14] These methods range from MD simulations with classical force fields to accurate quantum mechanics (QM) calculations to understand specific local interactions in detail. The use of classical force fields, such as OL15 or bsc,[15,16] for the study of DNA has demonstrated high precision and reliability[17] but they show limitations to deal with GQ since their parameters are usually optimized considering dDNA structures and not GQ. In this sense, optimization of force field (FF) parameters for the correct description of the GQ has been the subject of work by Sponer et al.[15]

Despite the complexity of biological systems, many phenomena may be studied just taking into account a relatively localized region. Assuming that the QM treatment is needed,

the choice is between a QM cluster approach or the use of the hybrid QM/MM approach.[18] Although the QM/MM approach has been extensively used to date,[19,20] it also presents shortcomings that may affect the quality of the results: *a*) The QM/MM method is based on partitioning the system of interest in a QM and a MM region, which introduces some arbitrariness in the calculation. *b*) The partition can lead to the cleavage of covalent bonds, and one needs an appropriate treatment of this boundary. *c*) Electrostatic interactions between the QM and MM regions may be considered at different levels. If one needs to overcome the inherent limitations of an artificial QM/MM partition or the property of interest cannot be localized in a small region, a full semi-empirical (SE) description of the system may be a good alternative.[21,22] SE methods fill the gap between MM and first-principles QM methods, being few orders of magnitude slower than MM methods but still orders of magnitude faster than first-principles QM methods. On the other hand, SE methods may provide reasonable accuracy in geometries and energies, although this is often system-dependent. In addition, SE methods can take polarization and charge transfer effects into account but they have usually problems describing dispersion and hydrogen bonds.[23] Nevertheless, more recently, corrections have been added to overcome these difficulties to improve their performance considerably.[24–28] That is, corrections were added by Rezac et al. in 2009 in one of the most popular Hamiltonians, PM6, to improve the description of non-covalent interactions in the modified PM6-DH Hamiltonian,[29] which later was revised in the PM6-DH2 Hamiltonian by Korth et al.[30–32] Recently, in 2016, an extensive review of the SE methods for non-covalent biochemical interactions have been carried out by Qiang Cui et al.[22]

The software improvements on theoretical methods and their optimization for parallelization make possible nowadays accurate *ab initio* calculations, in which the computational cost scales linearly with the system size, such as LS-DFT[33] or near linear-scaling Coupled-Cluster.[34] The treatment of the entire system with a first-principles quantum method ensures an accurate description of the interaction between ligands and biomolecules. Therefore, it can be of high relevance to study DNA-Ligand interactions.

In the present paper, we use a LS-DFT approach, namely, with the SIESTA[35] code, using the LMKLL density functional,[36] and the DZDP basis set,[36] to analyze the interaction of small chemical species with dDNA and GQ. The results are compared with the semi-empirical methods PM6-DH2[29,30] and PM7[37] (using the MOPAC[38] package), QM/MM calculations, near linear-scaling DLPNO-CCSD(T) calculations[39] (implemented in the ORCA *v4.2.1* package[40]), and benchmark database.[41] Modeling of GQ have been the subject of substantial interest recently.[14] As far as we know, this is the first time that LS-DFT methods have been used to study ligand-GQ interactions by taking into account the whole GQ structure of more than 1000 atoms without reduction to smaller models. We also demonstrate the adequacy of these LS-DFT approaches to treat DNA in its multiple forms and their interactions with small chemical species.

Methods

To perform SE geometry optimizations, we used MOPAC *v8.0.0* with the PM6-DH2 and PM7 Hamiltonians. The default Eigenvector Following routine (EF) was used,[42] whereas the default SCF criterion was changed to 1×10^{-8}. The minimum trust radius was set at 0.0001 Å/rad and a damping factor of 10 was added with the SHIFT method to improve the SCF procedure.[43] For the LS-DFT calculations, SIESTA *4.1-b3* software was used.[35] Geometry optimizations were performed with the LMKLL van der Waals functional,[36] which includes dispersion corrections, being highly appropriate for the characterization of weak forces within the ligand-DNA interaction. The modified Broyden algorithm was used for geometry optimization.[44] SCF convergence was accelerated with the Pulay method[45] keeping a history of 4 past density matrices, the density matrix mixing weight was set to 0.005. For the basis set, a 30 meV energy shift was used along with a 150 Ry mesh cut off for real space integration. We also used a SCF tolerance of 1×10^{-5} eV and we set the max force tolerance at 0.02 eV/Å for the dDNA and at 0.1 eV/Å in the case of GQ structures. In the case of GQ

systems the max force tolerance was established after doing some tests where it was observed that although the tolerance is quite high, it does not compromise either the total energy or the relaxed geometry, at the same time that it allows to reduce the computation time (see Table S6). An optimized double zeta plus double polarization (DZDP) basis set was used for each atom[46] along with Troullier-Martins norm-conserving pseudopotentials[47,48] that were generated with the ATOM package included in SIESTA software, whereas for the G-tetrads used for energetic calibration, the psml pseudopotentials[49] from the website www.pseudo-dojo.org were used. It must be mentioned that, in the SIESTA method the matrix elements are computed with linear-scaling algorithms while the diagonalization is proportional to $O(\mathrm{N}^3)$.

DLPNO-CCSD(T)[39] single-point calculations were also performed with the ORCA *4.2.1* software.[40] The Ahlrichs's def2-SVP basis set[50] was used for all the calculations with the corresponding auxiliary bases of Weigend for RI-J[51] and RIJCOSX approximations.[52]

We also run QM/MM geometry optimizations for the 1n37 and 2jwq PDB structures, that were performed at M11L/6-31+G(d,p):AMBER and B3LYP-D3(GD3BJ)/6-31+G(d,p):AMBER level, respectively, as implemented in Gaussian*16*.[53] Quadratic Convergence (QC)[54] SCF procedure was used with a maximum amount of 1500 cycles and the maximum size for an optimization step was changed from 30 to 1. In both cases the containing ligand was treated as the QM part and the rest of the molecule (the DNA) as the MM layer, no boundaries needed to be used, because the ligand and DNA are not linked.

Considered structures

DNA base pairs.

Initial geometries were taken from the benchmark database performed by Hobza and coworkers, which contains computationally optimized and experimental DNA base pair structures interacting through H-bonds and by $\pi - \pi$ stacking.[55] Interaction energies (ΔE_{int}) for these systems were computed by single-point calculations, at the different levels we compare in

this study, and subtracting to the total energy (E_{tot}) the energy of each DNA base fragment (E_{frag}) as shown in Eq 1:

$$\Delta E_{int} = E_{tot} - E_{frag1} - E_{frag2} \tag{1}$$

Intercalated Phenanthroline (phen) in DNA base pairs.

Optimized structures at M06-2X/6-31+G(d,p) level were taken from our previous work[56] and single-point calculations were performed at the different levels of calculation we compare in this work. The interaction energy for the phen/DNA systems was calculated as shown in Eq 2, where one fragment is composed of all DNA atoms and the second is the phenanthroline ligand:

$$\Delta E_{int} = E_{tot} - E_{DNA} - E_{phen} \tag{2}$$

G-tetrads.

We considered four model systems from the work of Fonseca Guerra and coworkers: G_4MG_4, aG_4MG_4, GQM, and $GQ_{4Na}M$, where M corresponds to the different metal cations (Li, Na, K, Rb, and Cs) placed in the ion channel[57] The first two models (G_4MG_4 and aG_4MG_4) only contain the metal cation and the guanine bases, whereas the other two models (GQM and $GQ_{4Na}M$) contain also the sugar and phosphate backbone or side loop, which is terminated by H^+ or Na^+, respectively, to compensate for the negative charge of the phosphate group.[58] For these systems, single-point calculations were performed at the different levels of calculation we compare in this work and the interaction energy was calculated following the same formula as in the original work, Eq 3:

$$\Delta E_{int} = E\left(G_4MG_4\right)_{gas} - E\left(G_4G_4\right)_{gas} - E\left(M\right)_{gas} \tag{3}$$

where G_4MG_4 is replaced by aG_4MG_4, GQM, or $GQ_{4Na}M$, and the empty scaffold

G_4G_4 is replaced by aG_4G_4, GQ, or GQ_{4Na}, which are the same structures without the metal cation.

1n37 and 2jwq systems

To perform larger geometry optimizations, two different DNA structures were taken from the PDB: 1) the 1n37 octamer,[59] which contains Respinomycin D ligand intercalated in a dDNA structure; and 2) the GQ structure 2jwq,[60] which has two MMQ-1 units, each bounded at the end-staking of both sides of the GQ. The systems were neutralized by adding an alkaline cation close to each phosphate group at a distance of 2.8 Å, Na^+ in the dDNA and K^+ for the GQ. In the case of the GQ, two additional K^+ were added, centered in the ion-channel between G-tetrads, resulting in a +2 charged system.

Results and Discussion

Geometrical discussion

Duplex DNA

Geometry optimizations were run departing from the PDB structures with SE, QM/MM, and LS-DFT methods. In Figure 1, we plot the superposition of the 1n37 structure from the PDB with the optimized geometries. SE results are the ones that shows the highest deviation from the initial reference structure giving the highest Root-Mean-Square Deviation (RMSD) values. Namely, PM6-DH2 shows a RMSD of 1.84 Å, although the structure seems to be maintained. However, PM7 gives a much higher RMSD, 3.02 Å, and the general structure is clearly not well superimposed. QM/MM method gives a slightly better result than PM6-DH2 with a 1.78 Å RMSD value and a lower deviation for the structure corresponding to the ligand than for the dDNA part. Finally, the LS-DFT method is by far the best method reproducing the PDB structure, giving an excellent RMSD value of 0.45 Å. Such results show that the SIESTA method and software not only has good performance in terms of computing

time for large systems of thousands of atoms, but also that the obtained structures are very accurate in terms of geometry.
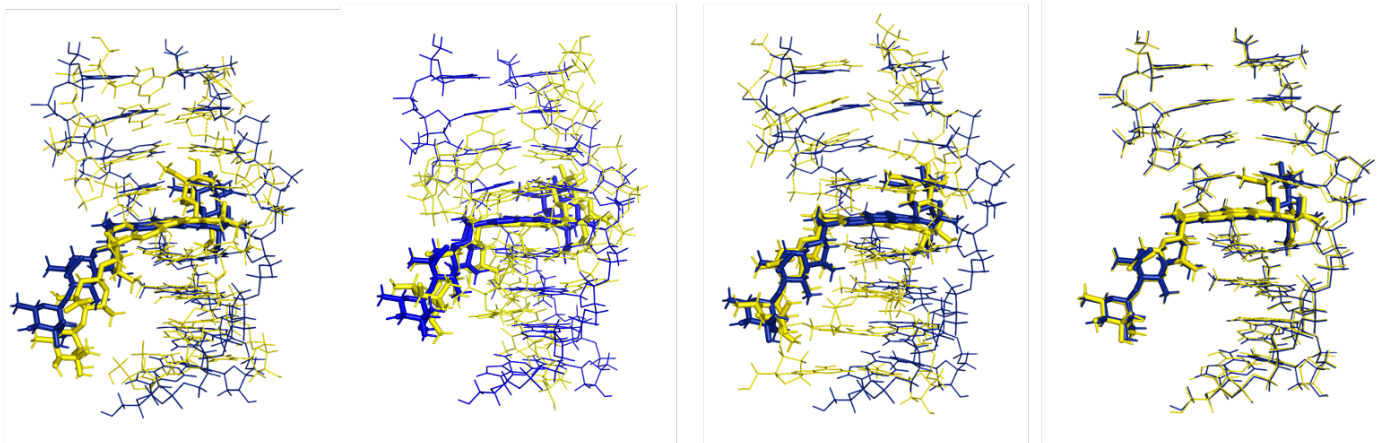

Figure 1: Overlap structures of 1n37 (in blue) with the optimized geometries (in yellow). From left to right with the methods PM6-DH2,PM7, M11L/6-31+G(d,p):AMBER and LMKLL/DZDP. RMSD values of 1.84 Å, 3.02 Å, 1.78 Å and 0.45 Å were obtained, respectively.

To further calibrate the performance of these methods for structural characterization, we also used other geometrical parameters characteristic of these systems with drugs intercalating between DNA base pairs. We analyzed the hydrogen bond lengths, the so-called twist angle (ϑ) and the rise (R) parameter, calculated in the same way as done in previous works[61] (see Supporting Information for definition), and we compared them to the counterparts of the original 1n37 PDB structure. Such values are depicted in Table 1.

First, it is worth mentioning that the PM7 semi-empirical method gives the worst geometrical results for the 1n37 system. Although the main structural features are kept, some of the base pairs, such as A_3-T_6 or T_8-A_1, are separated. The structure is not kept straight, in a conformation with all bases stacked, but curved in a "C" shape. For these reasons, we will not further consider this structure for deeper geometrical analysis.

Let us start analyzing the twist angle parameter, calculated in the same way as in previous works.[62] LMKLL is the method that gives the best results for the twist angle, with differences from 0.0º to 1.3º in most cases. Only two twist angles differ by more than 2º: the twist

10

Table 1: Hydrogen bond lengths, rise (R) parameter and twist angle (ϑ) of the original 1n37 PDB structure in parentheses, after LS-DFT optimization at LMKLL/D2DZ level with SIESTA in bold, after optimization with PM6-DH2 in italics, after optimization with PM7 in italics plus bold and finally, QM/MM optimization at M11-L/6-31+G(d,p):AMBER level in normal script (see Figure S1 of the Supporting Information to check the labels of the atoms corresponding to the hydrogen bonds).

Base Pairs	Purine···Pyrimidine	Distance (Å)	R (Å)	ϑ (º)
A_1-T_8	N_6···O_4	(2.93)/**3.16**/*2.78*/***2.98***/2.88		
	N_1···N_3	(2.84)/**2.85**/*3.12*/***3.15***/2.92		
	C_2···O_2	(3.47)/**3.38**/*4.38*/***3.97***/3.66		
			(2.85)/**3.61**/*3.44*/***3.73***/3.53	(23.9)/**23.2**/*27.2*/***8.0***/22.2
G_2-C_7	O_6···N_4	(2.83)/**2.92**/*2.87*/***3.02***/2.89		
	N_1···N_3	(2.90)/**2.95**/*2.86*/***2.93***/2.90		
	N_2···O_2	(2.84)/**2.93**/*2.83*/***2.83***/2.78		
			(3.60)/**3.15**/*3.47*/***1.42***/2.90	(23.3)/**21.8**/*26.2*/***-43.4***/30.5
A_3-T_6	N_6···O_4	(2.87)/**3.11**/*2.85*/***5.61***/3.02		
	N_1···N_3	(2.94)/**3.00**/*3.00*/***7.33***/2.87		
	C_2···O_2	(3.70)/**3.84**/*4.00*/***8.84***/3.45		
			(3.54)/**3.46**/*3.54*/***2.89***/3.57	(21.6)/**21.6**/*20.5*/***15.3***/20.9
C_4-G_5	O_6···N_4	(2.92)/**3.80**/*2.87*/***2.89***/2.95		
	N_1···N_3	(2.91)/**3.29**/*2.84*/***3.00***/2.91		
	N_2···O_2	(2.80)/**2.87**/*2.82*/***2.92***/2.78		
			(5.66)/**5.75**/*5.94*/***5.51***/6.05	(5.2)/**24.5**/*30.5*/***15.4***/21.9
G_5-C_4	O_6···N_4	(2.87)/**2.93**/*2.85*/***2.88***/2.84		
	N_1···N_3	(2.88)/**2.99**/*2.87*/***2.91***/2.90		
	N_2···O_2	(2.83)/**2.96**/*2.83*/***2,86***/2.84		
			(3.52)/**3.31**/*3.16*/***3.10***/3.35	(25.0)/**22.7**/*26.2*/***28.4***/23.6
T_6-A_3	N_6···O_4	(2.96)/**3.27**/*2.84*/***3.25***/2.89		
	N_1···N_3	(2.89)/**3.04**/*3.05*/***2.89***/2.86		
	C_2···O_2	(3.61)/**3.79**/*4.18*/***3.35***/3.46		
			(3.15)/**3.07**/*3.31*/***2.90***/4.02	(26.8)/**26.5**/*26.2*/***20.2***/29.7
C_7-G_2	O_6···N_4	(2.80)/**2.88**/*2.83*/***2.94***/2.87		
	N_1···N_3	(2.94)/**3.01**/*2.85*/***2.98***/2.85		
	N_2···O_2	(2.99)/**3.08**/*2.86*/***2.97***/2.89		
			(3.18)/**3.14**/*2.73*/***0.44***/2.79	(0.7)/**-0.9**/*-12.5*/***18.8***/-17.0
T_8-A_1	N_6···O_4	(7.40)/**7.86**/*3.04*/***5.41***/3.46		
	N_1···N_3	(7.31)/**7.89**/*4.67*/***6.08***/7.93		
	C_2···O_2	(6.12)/**6.68**/*7.37*/***7.57***/11.19		

angle formed between C_4-G_5 and G_5-C_4 and that formed between G_5-C_4 and T_6-A_3. On the other hand, it is unclear which approach gives the worst results for the twist angle. In some cases, QM/MM yields the worst results, whereas in others, PM6-DH2 does. In general, the results for the twist angle given by the three different approaches agree with the analysis of the RMSD.

Regarding the Rise parameter (R), the QM/MM approach at M11-L/6-31+G(d,p): AMBER level of theory gives the worst results with differences to the PDB structure higher than 0.5 Å in some cases. On the other hand, as a general trend, the best results are again given by the LS-DFT. Nevertheless, the SE approach also gives excellent results for the rise parameter (R), in some cases (A_1-T_8/G_2-C_7, G_2-C_7/A_3-T_6, and A_3-T_6/C_4-G_5), even better than the LS-DFT approach. Thus, it is not clear which of the two approaches, LS-DFT or semi-empirical, including dispersion yield the best results.

Interesting trends are also observed for hydrogen bond distances. In the case of the hydrogen bond distances of the A-T base pairs, we see that the SE approach gives the worst behavior. Indeed, there is a considerable lengthening, especially for the $N_1 \cdots H$-N_3 and C_2-$H \cdots O_2$ hydrogen bonds. LS-DFT approach gives the best results for the hydrogen bond distances again, although in general, they are elongated with respect to the PDB structure, especially for the $N_1 \cdots H$-N_3 and C_2-$H \cdots O_2$ hydrogen bonds. Nevertheless, it must be mentioned that the QM/MM approach at M11-L/6-31+G(d,p):AMBER also gives the best results in some of the N_6-$H \cdots O_4$ hydrogen bonds. On the other hand, there is a better agreement between theoretical hydrogen bond distances and those corresponding to the original PDB structure for the G-C base pairs, with LS-DFT giving the best results.

Summarizing, the semi-empirical approach with the PM6-DH2 Hamiltonian gives in some cases better results than the most popular QM/MM approach at M11-L/6-31+G(d,p):AMBER level of theory. On the other hand, the LS-DFT method gives an excellent agreement between the geometrical parameters of the PDB structure and those of the optimized structure. SIESTA software gives us excellent results for this biomolecular system of 665 atoms in a

reasonable computation time.

G-quadruplex

We carried out an analysis for the systems based on GQ by using similar structural parameters as for the dDNA (See Table 2): RMSD from the PDB reference structure, hydrogen bond lengths, rise distances, and twist angle, in this case as defined by Chung et al.[63](see Supporting Information for the nomenclature in Figure S3, Rise definition, and twist angle definition in Figure S2). The 2jwq PDB structure for the GQ was optimized at PM6-DH2 and PM7 level (in the case of the semi-empirical approach), at B3LYP-D3(GD3BJ)/6-31+G(d,p):AMBER level (in the case of the QM/MM approach) and at LMKLL/DZDP level (an LS-DFT method), see Figure 2. Semi-empirical methods yielded higher RMSD values than the LS-DFT method. The PM6-DH2 and PM7 methods gave RMSD values of 0.98 and 2.06 Å, respectively. Quite interestingly, the RMSD values obtained with the semi-empirical models for the systems based on GQ were lower than those obtained for dDNA, even though the GQ used in this work have approximately twice as many atoms as the studied dDNA systems. This better agreement could be due to a more rigid structure of the GQ because the stacking of the G-tetrads creates a more rigid structure, which in turn increase the rigidity of the side loops.
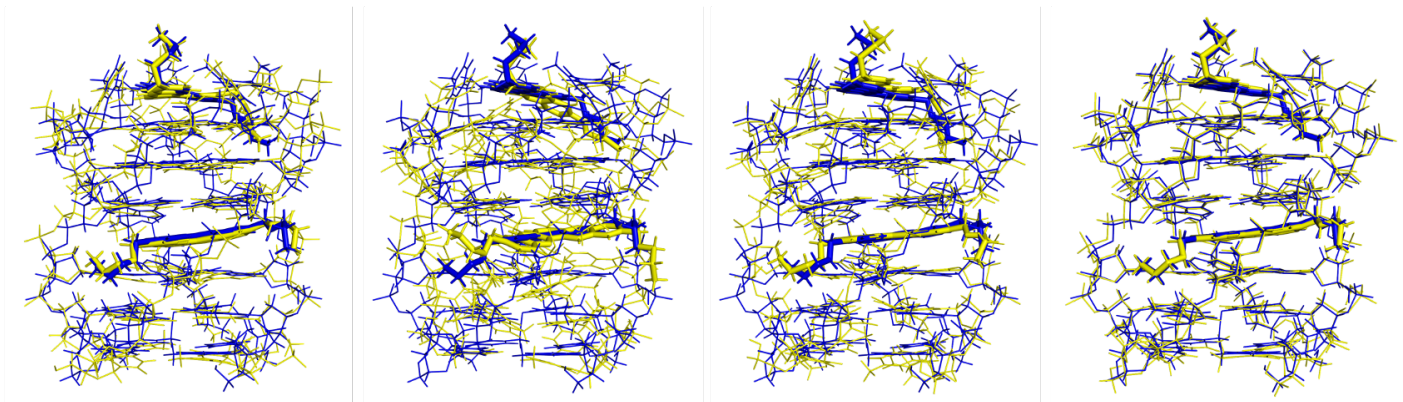

Figure 2: Overlapping of 2jwq PDB structure (in blue) with the optimized geometries (in yellow) with the PM6-DH2, PM7, B3LYP-D3(GD3BJ)/6-31+G(d,p):AMBER and LMKLL methods, from left to right. RMSD values of 0.98 Å, 2.06 Å, 1.05 Å and 0.24 Å, respectively.

Table 2: Hydrogen bond lengths, rise (R) parameter and twist angle (ϑ) of the original 2jwq PDB structure in parentheses, after LS-DFT optimization at LMKLL/D2DZ level with SIESTA in bold, after optimization with PM6-DH2 in italics, after optimization with PM7 semi-empirical method in italic plus bold, and at the B3LYP-D3(GD3BJ)/6-31+G(d,p):AMBER level in normal script (see Figure S3 of the Supporting Information to check the labels of the atoms corresponding to the hydrogen bonds and the labels for guanines in the tetrads).

Tetrads	Guanine···Guanine	Distance (Å)	R (Å)	ϑ (º)
	G_4O_6···$G_{18}N_1$	(3.01)/**2.90**/*2.90*/***3.00***/2.92		
	G_4N_7···$G_{18}N_2$	(2.88)/**3.00**/*3.04*/***3.01***/2.95		
	$G_{18}O_6$···$G_{11}N_1$	(3.14)/**2.94**/*2.88*/***2.96***/2.94		
	$G_{18}N_7$···$G_{11}N_2$	(2.77)/**2.93**/*3.00*/***3.07***/2.88		
$G_{4,11,18,25}$	$G_{11}O_6$···$G_{25}N_1$	(2.85)/**2.90**/*2.90*/***2.81***/2.89		
	$G_{11}N_7$···$G_{25}N_2$	(3.41)/**3.16**/*3.06*/***3.15***/3.01		
	$G_{25}O_6$···G_4N_1	(2.92)/**2.91**/*2.82*/***2.88***/2.89		
	$G_{25}N_7$···G_4N_2	(2.93)/**3.02**/*2.97*/***3.06***/2.88		
			(3.38)/**3.41**/*3.31*/***3.33***/3.39	(28.7)/**25.3**/*26.8*/***25.3***/24.8
	G_5O_6···$G_{19}N_1$	(3.10)/**2.91**/*2.74*/***2.82***/3.17		
	G_5N_7···$G_{19}N_2$	(3.45)/**2.98**/*3.20*/***2.96***/2.92		
	$G_{19}O_6$···$G_{12}N_1$	(2.53)/**3.08**/*2.86*/***2.87***/2.98		
	$G_{19}N_7$···$G_{12}N_2$	(2.43)/**2.98**/*3.14*/***3.01***/2.86		
$G_{5,12,19,26}$	$G_{12}O_6$···$G_{26}N_1$	(2.91)/**2.85**/*2.79*/***2.98***/2.95		
	$G_{12}N_7$···$G_{26}N_2$	(3.77)/**3.38**/*3.53*/***2.96***/3.13		
	$G_{26}O_6$···G_5N_1	(3.08)/**2.92**/*2.80*/***2.89***/2.90		
	$G_{26}N_7$···G_5N_2	(3.17)/**3.00**/*3.06*/***2.89***/2.96		
			(3.19)/**3.26**/*3.17*/***3.21***/3.42	(16.4)/**21.9**/*19.5*/***30.3***/28.5
	G_6O_6···$G_{20}N_1$	(2.83)/**2.85**/*2.77*/***2.88***/2.75		
	G_6N_7···$G_{20}N_2$	(2.85)/**3.02**/*3.06*/***2.93***/3.02		
	$G_{20}O_6$···$G_{13}N_1$	(2.99)/**2.99**/*2.80*/***2.87***/2.82		
	$G_{20}N_7$···$G_{13}N_2$	(2.80)/**3.00**/*3.18*/***3.05***/2.89		
$G_{6,13,20,27}$	$G_{13}O_6$···$G_{27}N_1$	(2.89)/**2.99**/*2.83*/***2.87***/2.87		
	$G_{13}N_7$···$G_{27}N_2$	(3.05)/**3.04**/*3.17*/***2.98***/2.84		
	$G_{27}O_6$···G_6N_1	(2.86)/**2.86**/*2.80*/***2.93***/2.85		
	$G_{27}N_7$···G_6N_2	(2.91)/**3.04**/*3.02*/***2.98***/2.86		

The LMKLL linear-scaling DFT method gave the best agreement to the experimental structure again, with a RMSD of 0.24 Å. On the other hand, QM/MM and PM6-DH2 show a similar RMSD, 1.05 Å for the former and 0.98 for the latter, whereas PM7 gives the worst result, namely, a RMSD of 2.06 Å. There is a tendency for the different computational methods to reduce the range in distances between the heavy atoms in the hydrogen bonds with respect to the values observed in the PDB. The PDB structure shows a range in $O_6 \cdots N_1$ distances between 2.53-3.14 Å, whereas for LMKLL, QM/MM, PM6-DH2, and PM7, the ranges are 2.85-3.08 Å, 2.75-3.17 Å, 2.74-2.9 Å, and 2.81-3.00 Å, respectively. The differences are even larger in the case of the $N_7 \cdots N_2$ distances, in which the PDB range is 2.43-3.77 Å while for the LMKLL, QM/MM, PM6-DH2, and PM7 are 2.93-3.38 Å, 2.82-3.13 Å, 2.97-3.53 Å and 2.89-3.15 Å, respectively. The LMKLL is again the method that better reproduces these distances, followed by QM/MM, PM6-DH2, and PM7 methods. These changes in distance, although small, cause the guanine bases to rotate directly affecting the twist angle. To calculate the twist angle we used a method devised by Phan et al.[63] The rotation angle between the G-tetrads was established as the average value obtained by calculating the angle of two guanines stacked between the vectors formed for each guanine by the coordinate of the C_8 atom and the midpoint of the coordinates of the N_1 and C_2 atoms. We have found that the twist angle for $G_{4,11,18,25}$ - $G_{5,12,19,26}$ is well reproduced by all methods. However the twist angle for $G_{5,12,19,26}$ - $G_{6,13,20,27}$ (16.4º) is reproduced appropriately by LMKLL (21.9º) and QM/MM (19.5º), whereas PM6-DH2 (30.3º) and PM7 (28.5º) semi-empirical methods show larger discrepancies. Finally, the Rise distance is well reproduced by all methods with differences with respect to the PDB reference structure within 0.2 Åin all cases.

In summary, the LS-DFT at the LMKLL/DZDP level of theory is the method that best describes the geometrical structure of the GQ system. In general, we observe a better agreement between the optimized and PDB structures for GQ than for dDNA, which can be related to the more rigid structure of the former. The geometrical results obtained for 1n37 dDNA and 2jwq GQ confirm that our LS-DFT approach is appropriate to optimize DNA-

type biomolecules. On the other hand, PM6-DH2 describes reasonably well both dDNA and GQ DNA systems but not the PM7 method. In addition, the PM6-DH2 approach may yield geometries of similar quality as for the popular QM/MM methodology. However, the PM7 Hamiltonian shows the most significant deviations in geometries. Finally, it must be said that the popular QM/MM approach we used here at B3LYP-D3(GD3BJ)/6-31+G(d,p):AMBER level may also lead to qualitatively correct structures.

Energetics discussion

DNA base pairs

Fig. 3 shows graphically the trends for the interaction energies (ΔE_{int}) of different DNA hydrogen-bonded and stacked base-pairs in the gas phase for the different computational methods studied in our work. The corresponding values are depicted in Table S1.

It is observed that the DLPNO-CCSD(T) and LMKLL methods reproduced the interaction energies of the benchmark references[55] accurately. The obtained ΔE_{int} for LMKLL and DLPNO-CCSD(T) are virtually identical, and both methods reproduce the reference interaction energies. The correlation coefficient r^2 with respect to the reference data is 0.996 for both methods, denoting a very good performance to describe the trends in interaction energies of these systems. The mean absolute error (MAE) for DLPNO-CCSD(T) is 2.1/1.8 kcal/mol for hydrogen-bonded/stacked base pairs, whereas the MAE for LMKLL is only 1.1/2.1 kcal/mol, respectively. In the case of systems where H-bonding links the bases, the interaction energies obtained by LS-DFT are systematically higher in absolute value than those obtained with DLPNO-CCSD(T). In contrast, for the stacked base pair systems, the ΔE_{int} are similar in all cases. On the other hand, the LS-DFT H-bonded systems are slightly more accurately described than stacked ones. This fact could be due to the use of the LMKLL functional, which improves the description of non-covalent interactions, but especially when hydrogen bonding plays an important role.[36] On the other hand, the SE methods considerably underestimate the ΔE_{int} for these systems, with a MAE of 9.2 and 11.6 kcal/mol for
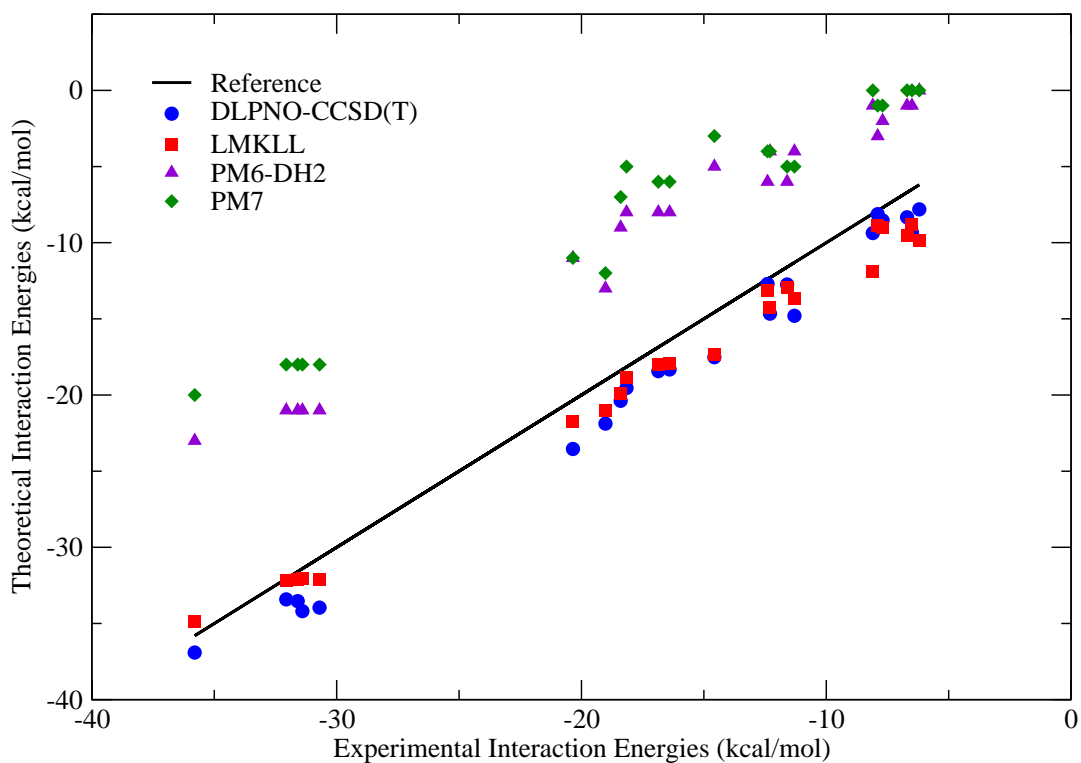
16

Figure 3: Interaction energies (kcal/mol) for DNA base pair systems, with CCSD energies from reference 38 as the reference data. The correlation coefficients r^2 for the DLPNO-CCSD(T), LMKLL, PM6-DH2, and PM7 are 0.996, 0.996, 0.990, 0.984, respectively.

PM6-DH2 and PM7, respectively, in the case of H-bonded base pairs and 6.1 kcal/mol and 7.4 kcal/mol, respectively, for stacked base pairs. Nevertheless, all methods show a similar r^2 correlation with respect to the reference data, indicating that even semi-empirical models may describe qualitatively the trends in interaction energies.

Intercalated Phenanthroline in DNA base pairs (phen/DNA)

To validate the adequacy of these methods to treat the interaction of ligands intercalated between DNA base pairs through weak interactions, we used a previously characterized system,[56] where a phenanthroline ligand is intercalated between two pairs of bases through both major groove (MG) and minor groove (mg). In this previous work, the MP2/6-31G*(0.25) theory level was used, based on the correction performed by Reha et al., which considers a modification of d-polarized basis function,[64] leading to a better agreement with CCSD(T) benchmark energies. We performed calculations for the same systems in this work, taking the DLPNO-CCSD(T) method as a reference. We computed the ΔE_{int} between the ligand (phen) and the DNA base pairs. This ΔE_{int} was calculated by subtracting from the total energy the energy of the separated fragments: phen ligand and the DNA fragment. The trends obtained for the interaction energies are presented in Figure 4, whereas the values are depicted in Table S2 of the ESI.

The agreement between LS-DFT and the reference DLPNO-CCSD(T) results is outstanding. The difference in interaction energies is lower than 1 kcal/mol in most of the cases, whereas the MAE is only 0.9 kcal/mol. It must be mentioned that the LMKLL functional has described with greater accuracy the ΔE_{int} of the systems containing guanine and cytosine. Moreover, the LMKLL functional describes the energetic trends very accurately and give more negative interaction energies for the MG structures than for the mg ones.

On the other hand, we observe that the semi-empirical methods show a poor performance with very low interaction energies, namely ΔE_{int}, which are around 30 kcal/mol smaller in absolute value than those of DLPNO-CCSD(T). The MAE for both methods is very
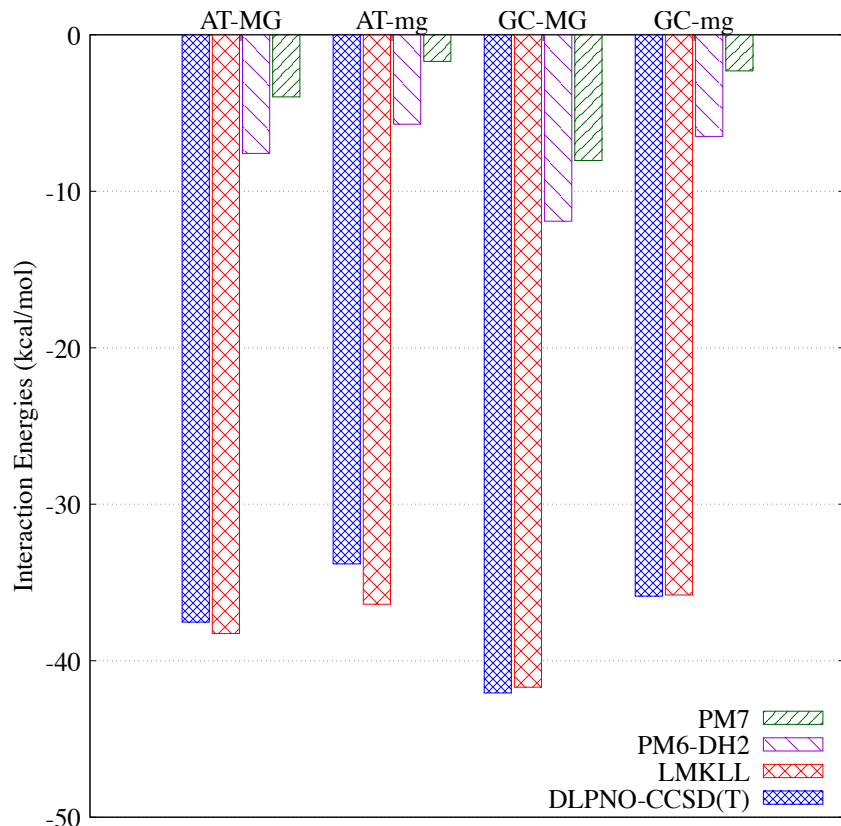
Figure 4: Interaction energies (kcal/mol) for the intercalated phenanthroline in DNA base pairs for the DLPNO-CCSD(T), LMKLL, PM6-DH2 and PM7 methods.

significant, 29.4 and 33.3 kcal/mol for PM6-DH2 and PM7, respectively. Nevertheless, it must be said that the trends in energetics, in which MG systems are more stable than mg ones, were correctly described by SE methods. For both PM6-DH2 and PM7 methods, the G-C/phen/C-G structure intercalating via MG was described correctly as the most stable system with a ΔE_{int} of -11.9 and -8.0 kcal/mol, respectively.

Summarizing, the description of the intercalation of ligands with DNA through weak interactions requires a reliable method to describe non-covalent interactions. Our results point to LS-DFT with the LMKLL functional as a method with an excellent performance in ΔE_{int} for ligand-DNA systems.

G-quadruplex structures

The obtained interaction energies for the systems based on G-tetrads according to Eq. 3 are represented graphically in Fig. 5 and their values depicted in Table S5 of the ESI.
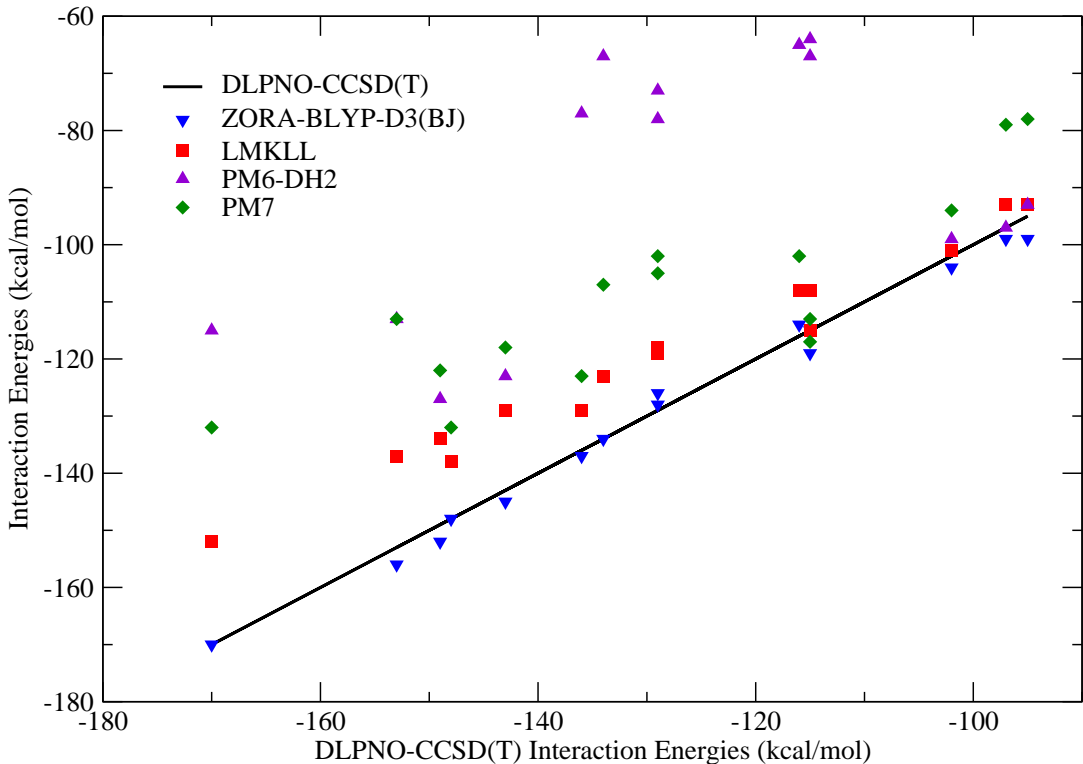

Figure 5: Interaction energies (kcal/mol) for the studied structures based on G-tetrads. r^2 correlation coefficients with respect to the DLPNO-CCSD(T) reference data for ZORA-BLYP-B3(BJ)[57] LMKLL, PM6-DH2, and PM7 are 0.993, 0.989, 0.509, 0.777, respectively.

LS-DFT agree not only with the DLPNO-CCSD(T) highly correlated benchmark calculations but also with the DFT calculations found in the bibliography for the same systems.[57] Our LS-DFT results tend to give smaller interaction energies than those found in the literature but the trend is described nicely with a correlation coefficient of 0.989 with respect to the benchmark DLPNO-CCSD(T) calculations. As in the case of the work of Fonseca-Guerra

et al.,[57] the Na metallic cation provides a significant stabilization for the G-tetrad structure and this stabilization decreases as the size of the metallic cation increases, (see Fig. 6). On the other hand, the semi-empirical methods have a very poor behavior both qualitatively and quantitatively. There are significant MAEs for the interaction energies, namely 22.5 and 40.9 kcal/mol for PM7 and PM6-DH2, respectively. Moreover, the poor r^2 correlation coefficients, 0.509 for PM6-DH2 and 0.777 for PM7, reveal high limitations of semi-empirical methods to reproduce the correct trends in stabilization energies. As observed in Fig. 6, the semi-empirical methods show a less clear trend between stabilization of G-tetrads and metal size compared to those observed for DLPNO-CCSD(T) benchmark and LMKLL.
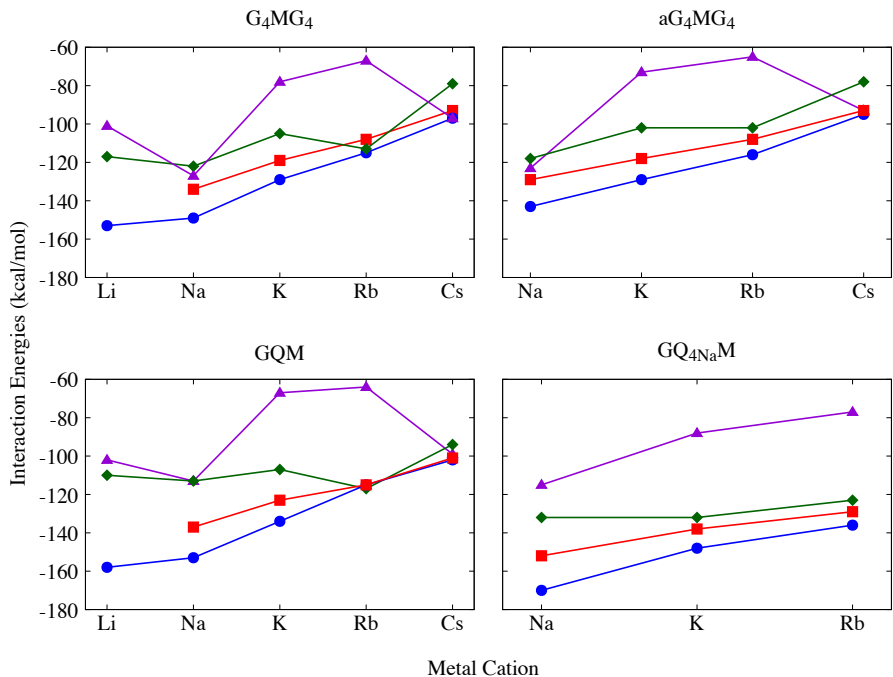

Figure 6: Interaction energies (kcal/mol) for the different G-tetrads studied systems (G_4MG_4 and aG_4MG_4 on top, and GQM and GQ_4NaM at bottom) with metal cations (Li, Na, K, Rb, and Cs) for each computational method used in this work. PM6-DH2 in purple, PM7 in green, LMKLL in red, and DLPNO-CCSD(T) in blue.

To sum up, considering all the interaction energies obtained for the different DNA systems, we can conclude that LS-DFT with the LMKLL functional can describe the weak interactions present in DNA systems. For G-tetrads, the ΔE_{int} deviate slightly from the reference ones, but the energetic trends are in good agreement with the reference calculations.

However, semi-empirical methods are very limited in the description of interaction energies. Therefore, the use of LS-DFT with the LMKLL functional including van der Waals corrections can be a very suitable strategy to analyze this kind of biological systems, in which a delicate balance of different non-covalent weak interactions is found.

Conclusions

Over the years, the computational power, the capacity for parallelization, and improvements in software have allowed the study of large biological systems using QM methods. In this sense, approaches based on LS-DFT are gaining in speed and efficiency, and they allow a more accurate description of the electronic structure of large biological systems. In this work we have studied three different DNA systems: i) DNA base-pairing, ii) models of duplex DNA interacting with a phen ligand and iii) G-tetrads stabilized with various alkaline metals. We have used a LS-DFT method (LMKLL/DZDP) and we have compared its performance with two semi-empirical methods incorporating dispersion corrections, PM6-DH2 and PM7, and QM/MM methods at B3LYP-D3(GD3BJ)/6-31+G(d,p):AMBER and at M11L/6-31+G(d,p):AMBER levels of theory. We have shown how the LMKLL applied through SIESTA reliably predicts both the geometries and the interaction energies for all these DNA systems, using experimental values and DLPNO-CCSD(T) benchmark calculations. On the other hand, the PM6-DH2 semi-empirical method has correctly described the geometries of these systems but not the PM7. However, both semi-empirical approaches are very limited in describing the interaction energies of these systems, with a degree of performance that is system-dependent. The present work opens the door for the computational investigation of large DNA systems using LS-DFT methods, which is particularly interesting for GQ, in which a proper balance of different weak non-covalent interactions, metal-ligand interactions, polarization and charge transfer is needed to give a realistic description of the system. In addition, it must be said that after the consolidation of conventional DFT

approaches, different works were reported in which a G2 modified composite methodology where the MP2 geometries and HF frequencies were substituted by the DFT ones and the QCISD(T) computations were replaced by CCSD(T) ones.[65][66] We already carried out similar computations by taking into account single-point calculations by means of CCSD(T) on DFT optimized geometries along with thermodynamic corrections to energy also at DFT level with systems of few tens of atoms.[67][68][69][70] In the forthcoming years, we believe that because of the developments in innovative algorithms, software, and hardware, some kind of LS-composite methodology could be feasible by means of single-point near LS-CCSD(T) energies on LS-DFT optimized structures by taking into account biological and chemical systems with thousands of atoms.

Acknowledgement

Supporting Information Available

Supporting Information includes definitions of the geometrical parameters for the duplex DNA and G-quadruplexes, and tables and graphics with the obtained interaction energies for the DNA base pairs, phen/DNA intercalation interaction and G-tetrads with the different used methods.

References

(1) Neidle, S. *Cancer drug design and discovery*; Elsevier, 2011.

(2) Rosenberg, B.; Vancamp, L.; Trosko, J. E.; Mansour, V. H. Platinum compounds: a new class of potent antitumour agents. *Nature* **1969**, *222*, 385–386.

(3) Braun, M. S.; Seymour, M. T. Balancing the efficacy and toxicity of chemotherapy in colorectal cancer. *Therapeutic advances in medical oncology* **2011**, *3*, 43–52.

(4) Chua, W.; Kho, P. S.; Moore, M. M.; Charles, K. A.; Clarke, S. J. Clinical, laboratory and molecular factors predicting chemotherapy efficacy and toxicity in colorectal cancer. *Critical reviews in oncology/hematology* **2011**, *79*, 224–250.

(5) Denny, W. A. DNA minor groove alkylating agents. *Current medicinal chemistry* **2001**, *8*, 533–544.

(6) Baraldi, P. G.; Bovero, A.; Fruttarolo, F.; Preti, D.; Tabrizi, M. A.; Pavani, M. G.; Romagnoli, R. DNA minor groove binders as potential antitumor and antimicrobial agents. *Medicinal research reviews* **2004**, *24*, 475–528.

(7) Nelson, S. M.; Ferguson, L. R.; Denny, W. A. Non-covalent ligand/DNA interactions: minor groove binding agents. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **2007**, *623*, 24–40.

(8) Hamilton, P. L.; Arya, D. P. Natural product DNA major groove binders. *Natural product reports* **2012**, *29*, 134–143.

(9) Arola, A.; Vilar, R. Stabilisation of G-quadruplex DNA by small molecules. *Current Topics in Medicinal Chemistry* **2008**, *8*, 1405–1415.

(10) Zhang, J.; Zhang, F.; Li, H.; Liu, C.; Xia, J.; Ma, L.; Chu, W.; Zhang, Z.; Chen, C.; Li, S., et al. Recent progress and future potential for metal complexes as anticancer drugs targeting G-quadruplex DNA. *Current medicinal chemistry* **2012**, *19*, 2957–2975.

(11) Balasubramanian, S.; Hurley, L. H.; Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nature reviews Drug discovery* **2011**, *10*, 261–275.

(12) Burger, A. M.; Dai, F.; Schultes, C. M.; Reszka, A. P.; Moore, M. J.; Double, J. A.; Neidle, S. The G-quadruplex-interactive molecule BRACO-19 inhibits tumor growth, consistent with telomere targeting and interference with telomerase function. *Cancer research* **2005**, *65*, 1489–1496.

(13) Mikami-Terao, Y.; Akiyama, M.; Yuza, Y.; Yanagisawa, T.; Yamada, O.; Yamada, H. Antitumor activity of G-quadruplex–interactive agent TMPyP4 in K562 leukemic cells. *Cancer letters* **2008**, *261*, 226–234.

(14) Ortiz de Luzuriaga, I.; Lopez, X.; Gil, A. Learning to Model G-Quadruplexes: Current Methods and Perspectives. *Annual Review of Biophysics* **2021**, *50*, 209–243.

(15) Zgarbová, M.; Sponer, J.; Otyepka, M.; Cheatham III, T. E.; Galindo-Murillo, R.; Jurecka, P. Refinement of the sugar–phosphate backbone torsion beta for AMBER

force fields improves the description of Z-and B-DNA. *Journal of chemical theory and computation* **2015**, *11*, 5723–5736.

(16) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A., et al. Parmbsc1: a refined force field for DNA simulations. *Nature methods* **2016**, *13*, 55.

(17) Galindo-Murillo, R.; Robertson, J. C.; Zgarbova, M.; Sponer, J.; Otyepka, M.; Jurecka, P.; Cheatham III, T. E. Assessing the current state of amber force field modifications for DNA. *Journal of chemical theory and computation* **2016**, *12*, 4114–4127.

(18) Banáš, P.; Jurečka, P.; Walter, N. G.; Šponer, J.; Otyepka, M. Theoretical studies of RNA catalysis: hybrid QM/MM methods and their comparison with MD and QM. *Methods* **2009**, *49*, 202–216.

(19) Lin, H.; Truhlar, D. G. QM/MM: what have we learned, where are we, and where do we go from here? *Theoretical Chemistry Accounts* **2007**, *117*, 185–199.

(20) Magalhães, R. P.; Fernandes, H. S.; Sousa, S. F. Modelling enzymatic mechanisms with QM/MM approaches: current status and future challenges. *Israel Journal of Chemistry* **2020**, *60*, 655–666.

(21) Hostaš, J.; Řezáč, J.; Hobza, P. On the performance of the semiempirical quantum mechanical PM6 and PM7 methods for noncovalent interactions. *Chemical Physics Letters* **2013**, *568*, 161–166.

(22) Christensen, A. S.; Kubar, T.; Cui, Q.; Elstner, M. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chemical reviews* **2016**, *116*, 5301–5337.

(23) Dannenberg, J. Hydrogen bonds: a comparison of semiempirical and ab initio treatments. *Journal of Molecular Structure: THEOCHEM* **1997**, *401*, 279–286.

(24) Thiriot, E.; Monard, G. Combining a genetic algorithm with a linear scaling semiempirical method for protein–ligand docking. *Journal of Molecular Structure: THEOCHEM* **2009**, *898*, 31–41.

(25) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz Jr, K. M. Formal estimation of errors in computed absolute interaction energies of protein- ligand complexes. *Journal of chemical theory and computation* **2011**, *7*, 790–797.

(26) Yang, Y.; Yu, H.; York, D.; Elstner, M.; Cui, Q. Description of phosphate hydrolysis reactions with the self-consistent-charge density-functional-tight-binding (SCC-DFTB) theory. 1. Parameterization. *Journal of chemical theory and computation* **2008**, *4*, 2067–2084.

(27) Tirado-Rives, J.; Jorgensen, W. L. Performance of B3LYP density functional methods for a large set of organic molecules. *Journal of Chemical Theory and Computation* **2008**, *4*, 297–306.

(28) Korth, M. Empirical Hydrogen-Bond Potential Functions—An Old Hat Reconditioned. *ChemPhysChem* **2011**, *12*, 3131–3142.

(29) Rezac, J.; Fanfrlik, J.; Salahub, D.; Hobza, P. Semiempirical quantum chemical PM6 method augmented by dispersion and H-bonding correction terms reliably describes various types of noncovalent complexes. *Journal of Chemical Theory and Computation* **2009**, *5*, 1749–1760.

(30) Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. A transferable H-bonding correction for semiempirical quantum-chemical methods. *Journal of chemical theory and computation* **2010**, *6*, 344–352.

(31) Yilmazer, N. D.; Korth, M. Comparison of molecular mechanics, semi-empirical quan-

tum mechanical, and density functional theory methods for scoring protein–ligand interactions. *The Journal of Physical Chemistry B* **2013**, *117*, 8075–8084.

(32) Yilmazer, N. D.; Heitel, P.; Schwabe, T.; Korth, M. Benchmark of electronic structure methods for protein–ligand interactions based on high-level reference data. *Journal of Theoretical and Computational Chemistry* **2015**, *14*, 1540001.

(33) Bowler, D. R.; Miyazaki, T. O(N) Methods in electronic structure calculations. *Reports on Progress in Physics* **2012**, *75*, 036503.

(34) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *The Journal of chemical physics* **2013**, *138*, 034106.

(35) Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. The SIESTA method for ab initio order-N materials simulation. *Journal of Physics: Condensed Matter* **2002**, *14*, 2745.

(36) Lee, K.; Murray, É. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher-accuracy van der Waals density functional. *Physical Review B* **2010**, *82*, 081101.

(37) Stewart, J. J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of molecular modeling* **2013**, *19*, 1–32.

(38) James, J. Stewart, Stewart Computational Chemistry MOPAC2016. 2016.

(39) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *The Journal of chemical physics* **2013**, *139*, 134101.

(40) Neese, F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 73–78.

(41) Saitow, M.; Becker, U.; Riplinger, C.; Valeev, E. F.; Neese, F. A new near-linear scaling, efficient and accurate, open-shell domain-based local pair natural orbital coupled cluster singles and doubles theory. *The Journal of chemical physics* **2017**, *146*, 164105.

(42) Baker, J. An algorithm for the location of transition states. *Journal of Computational Chemistry* **1986**, *7*, 385–395.

(43) Mitin, A. V. The dynamic "level shift" method for improving the convergence of the SCF procedure. *Journal of Computational Chemistry* **1988**, *9*, 107–110.

(44) Johnson, D. D. Modified Broyden's method for accelerating convergence in self-consistent calculations. *Physical Review B* **1988**, *38*, 12807.

(45) Banerjee, A. S.; Suryanarayana, P.; Pask, J. E. Periodic Pulay method for robust and efficient convergence acceleration of self-consistent field iterations. *Chemical Physics Letters* **2016**, *647*, 31–35.

(46) Junquera, J.; Paz, Ó.; Sánchez-Portal, D.; Artacho, E. Numerical atomic orbitals for linear-scaling calculations. *Physical Review B* **2001**, *64*, 235111.

(47) Troullier, N.; Martins, J. L. Efficient pseudopotentials for plane-wave calculations. II. Operators for fast iterative diagonalization. *Physical Review B* **1991**, *43*, 8861.

(48) Kleinman, L.; Bylander, D. Efficacious form for model pseudopotentials. *Physical Review Letters* **1982**, *48*, 1425.

(49) García, A.; Verstraete, M. J.; Pouillon, Y.; Junquera, J. The psml format and library for norm-conserving pseudopotential data curation and interoperability. *Computer Physics Communications* **2018**, *227*, 51–71.

(50) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, *7*, 3297–3305.

(51) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Physical chemistry chemical physics* **2006**, *8*, 1057–1065.

(52) Kossmann, S.; Neese, F. Comparison of two efficient approximate Hartee–Fock approaches. *Chemical Physics Letters* **2009**, *481*, 240–243.

(53) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(54) Bacskay, G. B. A quadratically convergent Hartree—Fock (QC-SCF) method. Application to closed shell systems. *Chemical Physics* **1981**, *61*, 385–404.

(55) Jurečka, P.; Šponer, J.; Černỳ, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical Chemistry Chemical Physics* **2006**, *8*, 1985–1993.

(56) Gil, A.; Melle-Franco, M.; Branchadell, V.; Calhorda, M. J. How the Intercalation of Phenanthroline Affects the Structure, Energetics, and Bond Properties of DNA Base Pairs: Theoretical Study Applied to Adenine–Thymine and Guanine–Cytosine Tetramers. *Journal of chemical theory and computation* **2015**, *11*, 2714–2728.

(57) Zaccaria, F.; Paragi, G.; Guerra, C. F. The role of alkali metal cations in the stabilization of guanine quadruplexes: why K+ is the best. *Physical Chemistry Chemical Physics* **2016**, *18*, 20895–20904.

(58) Nieuwland, C.; Zaccaria, F.; Guerra, C. F. Understanding alkali metal cation affinities of multi-layer guanine quadruplex DNA. *Physical Chemistry Chemical Physics* **2020**, *22*, 21108–21118.

(59) Searle, M. S.; Maynard, A. J.; Williams, H. E. DNA recognition by the anthracycline antibiotic respinomycin D: NMR structure of the intercalation complex with d (AGACGTCT) 2. *Organic & biomolecular chemistry* **2003**, *1*, 60–66.

(60) Hounsou, C.; Guittat, L.; Monchaud, D.; Jourdan, M.; Saettel, N.; Mergny, J.-L.; Teulade-Fichou, M.-P. G-quadruplex recognition by quinacridines: a SAR, NMR, and biological study. *ChemMedChem: Chemistry Enabling Drug Discovery* **2007**, *2*, 655–666.

(61) Gil, A.; Branchadell, V.; Calhorda, M. A theoretical study of methylation and CH/π interactions in DNA intercalation: methylated 1, 10-phenanthroline in adenine–thymine base pairs. *RSC advances* **2016**, *6*, 85891–85902.

(62) Galliot, A.; Gil, A.; Calhorda, M. J. Effects of oxygenation on the intercalation of 1, 10-phenanthroline-5, 6/4, 7-dione between DNA base pairs: a computational study. *Physical Chemistry Chemical Physics* **2017**, *19*, 16638–16649.

(63) Chung, W. J.; Heddi, B.; Schmitt, E.; Lim, K. W.; Mechulam, Y.; Phan, A. T. Structure of a left-handed DNA G-quadruplex. *Proceedings of the National Academy of Sciences* **2015**, *112*, 2729–2733.

(64) Řeha, D.; Kabelác, M.; Ryjácek, F.; Šponer, J.; Šponer, J. E.; Elstner, M.; Suhai, S.; Hobza, P. Intercalators. 1. Nature of stacking interactions between intercalators (ethidium, daunomycin, ellipticine, and 4 ', 6-diaminide-2-phenylindole) and DNA base pairs. Ab initio quantum chemical, density functional theory, and empirical potential study. *Journal of the American Chemical Society* **2002**, *124*, 3366–3376.

(65) Mebel, A.; Morokuma, K.; Lin, M. Modification of the GAUSSIAN-2 theoretical model: The use of coupled-cluster energies, density-functional geometries, and frequencies. *Journal of Chemical Physics* **1995**, *103*, 7414.

(66) Bauschlicher, C.; Partridge, H. A modification of the Gaussian-2 approach using density functional theory. *Journal of Chemical Physics* **1995**, *103*, 1788.

(67) Gil, A.; Bertran, J.; Sodupe, M. Gas Phase Dissociation Energies of Saturated AHn·+ Radical Cations and AHn Neutrals (A = Li-F, Na-Cl): Dehydrogenation, Deprotona-

tion, and Formation of AHn-2+-H2 Complexes. *Journal of American Chemical Society* **2003**, *125*, 7461–7469.

(68) Gil, A.; Bertran, J.; Sodupe, M. Effects of ionization on N-glycylglycine peptide: Influence of intramolecular hydrogen bonds. *Journal of Chemical Physics* **2006**, *124*, 154306.

(69) Gil, A.; Simon, S.; Rodriguez-Santiago, L.; Bertran, J.; Sodupe, M. Influence of the side chain in the structure and fragmentation of amino acid radical cations. *Journal of Chemical Theory and Computation* **2007**, *3*, 2210–2220.

(70) Gil, A.; Sodupe, M.; Bertran, J. Influence of ionisation on the conformational preferences of peptide models. Ramachandran surfaces of N-formhyl-glycine amide and N-formyl-alanine amide radical cations. *Journal of Computational Chemistry* **2009**, *30*, 1771–1784.

Supporting Information for

Semi-empirical and Linear-Scaling DFT Methods to Characterize duplex DNA and G-quadruplexes in Presence of Interacting Small Molecules

Iker Ortiz de Luzuriaga[ab], Sawssen Elleuchi[c], Khaled Jarraya[c], Emilio Artacho[adef], Xabier Lopez[bf*], and Adrià Gil[aghi*]

[a] CICnanoGUNE BRTA, Tolosa Hiribidea 76, E-20018, Donostia - San Sebastian.

[b] Polimero eta Material Aurreratuak: Fisika, Kimika eta Teknologia, Kimika Fakultatea, Euskal Herriko Uniberstitatea, UPV/EHU, 20080 Donostia, Euskadi, Spain.

[c] Laboratoire de Chimie Inorganique, LR17ES07, Université de Sfax, Faculté de Sciences de Sfax, 3000 Sfax, Tunisia.

[d] Theory of Condensed Matter, Cavendish Laboratory, University of Cambridge, J. J. Thomson Ave., Cambridge CB3 0HE, United Kingdom

[e] Ikerbasque, Basque Foundation for Science, 48011 Bilbao, Spain

[f] Donostia International Physics Center, 20018 Donostia, Spain.

[g] ARAID Foundation, Zaragoza, Spain

[h] Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH) CSIC- Universidad de Zaragoza, C/ Pedro Cerbuna 12, 50009, Zaragoza, Spain.

[i] BioISI – Biosystems and Integrative Sciences Institute, Faculdade de Ciências,Universidade de Lisboa, Campo Grande, 1749-016, Lisboa, Portugal

This Supporting Information includes definitions of the geometrical parameters for the duplex DNA and G-quadruplexes, and tables and graphics with the obtained interaction energies for the DNA base pairs, phen/DNA intercalation interaction and G-tetrads with with the different used methods.

Definitions for the R and twist angle (ϑ) parameters for the duplex DNA systems.

We defined the xy plane by the two atoms forming the $N_1 \cdots N_3$ hydrogen bond and the third atom for the definition of the xy plane is the C_2 atom of adenine (adenine and thymine base pairs) or the C_2 atom of cytosine (guanine and cytosine base pairs), as shown in Figure S1. Then, we define the R mean distance between the two base pairs as the difference between the mean z value of the atoms of the upper base pair and the one of the atoms of the lower base pair. We also analyzed the ϑ twist angle that may be defined from the schemes in Figure S1. That is, the dashed line joining the C_8 atom of the purine base to the C_6 atom of the pyrimidine is the long base-pair axis and the ϑ angle is defined as the rotation of one base pair around the center of its C_6–C_8 axis. Because the base pairs are not strictly planar and parallel after optimization, the angle of the optimized systems is defined as the angle between the projections on the xy plane of the C_6–C_8 axis of each base pair. Thus, the ϑ twist angle would be the dihedral angle between the vector in C_6–C_8 direction of the i base pair and the vector in the C_6–C_8 direction of the i+1 base pair in any step of the DNA chain.
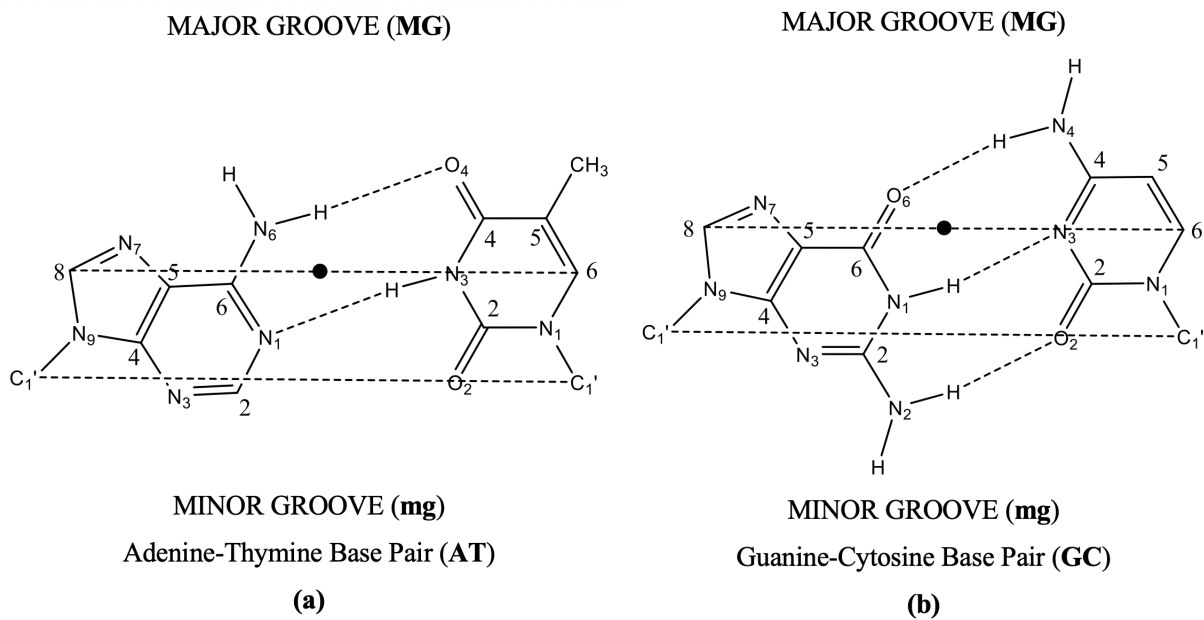
Figure S1: Scheme of the base pairs AT (a) and GC (b). The dashed line C_6–C_8 represents the long base-pair axis, which is roughly parallel to the C'_1–C'_1 line, where C'_1 stands for the sugar carbon atoms bonded to the bases. The twist angle (ϑ) is defined as the rotation around the midpoint of the C_6–C_8 axis (denoted by a dot).

Definitions for the R and twist angle (ϑ) parameters for the G-quadruplex systems.

We defined the xy plane by three guanine O_6 of the same G-tetrad. Then, we define the R mean distance between the two G-tetrads as the difference between the mean z value of the atoms of the upper G-tetrad and the one of the atoms of the lower G-tetrad. The ϑ twist angle, is defined as the angle between the lines formed by the guanine C_8 and the midpoint between N_1 and C_2, as can be seen in Figure S2.
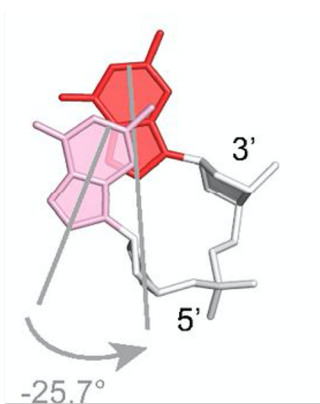
Figure S2: Representation of the stacking of two G-tetrad guanines.The ϑ twist angle, is defined as the angle between the lines formed by the guanine C_8 and the midpoint between N_2 and C_6.
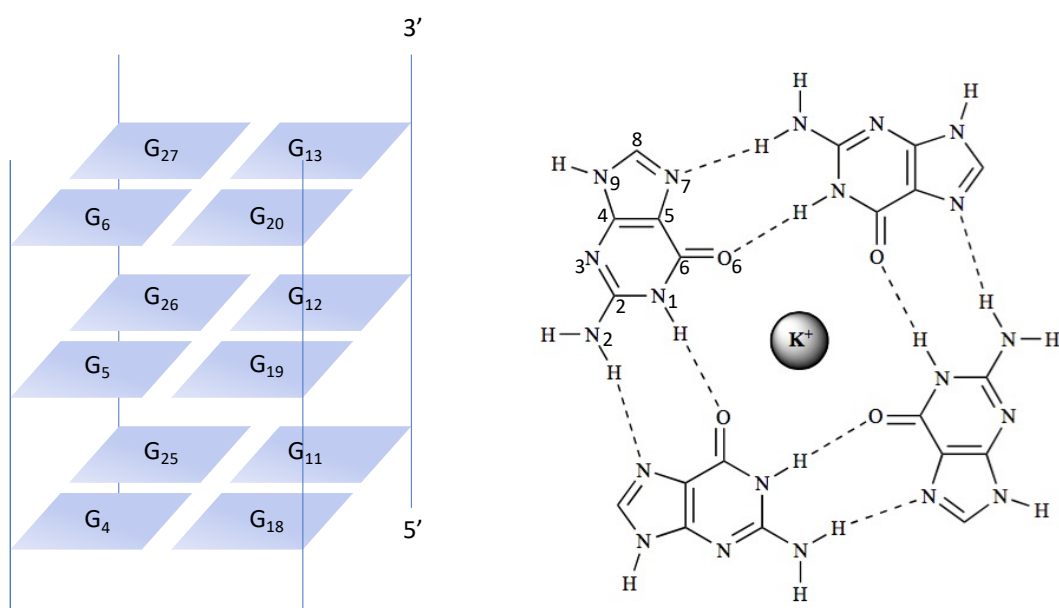
Figure S3: On the left, scheme of the stacking of three G-tetrads of guanines composing the 2JWQ PDB structure, and nomenclature for each guanine base. On the right, a G-tetrad, with the number used in the nomenclature.

Table S1: Interaction energies (kcal/mol) for different DNA base pairs. Abbreviations used in the first column: A, T, C, G – adenine, thymine, cytosine, guanine; m – methyl-; WC, H – Watson-Crick, Hoogsteen: OG, EG – optimized geometry, experimental geometry.

	System	Reference	DLPNO-CCSD(T)	LMKLL	PM6-DH2	PM7
H-bonded base pairs	G-C WC (OG)	-32.06	-33.42	-32.18	-18.06	-16.29
	mG-mC WC (OG)	-31.59	-33.53	-32.09	-17.78	-16.02
	A-T WC (OG)	-16.86	-18.43	-17.97	-8.60	-6.86
	mA-mT WC (OG)	-18.16	-19.54	-18.84	-9.09	-6.26
	A-T WC (EG)	-16.40	-18.32	-17.95	-8.46	-6.81
	G-C WC * (EG)	-35.80	-36.91	-34.84	-17.85	-16.03
	A-T WC (EG)	-18.40	-20.38	-19.93	-8.37	-6.99
	G-A HB (EG)	-11.30	-14.80	-13.67	-5.83	-6.16
	C-G WC (EG)	-30.70	-33.96	-32.11	-18.04	-16.45
	G-C WC (EG)	-31.40	-34.19	-32.02	-18.06	-16.47
	MAE	–	2.08	1.08	9.15	11.61
Stacked base pairs	G-C (OG)	-19.02	-21.88	-21.03	-10.99	-10.76
	mG-mC (OG)	-20.35	-23.54	-21.72	-10.28	-9.58
	A-T (OG)	-12.30	-14.66	-14.25	-5.22	-4.00
	mA-mT (OG)	-14.57	-17.52	-17.37	-6.14	-4.65
	A-T (EG)	-8.10	-9.37	-11.89	-3.13	-3.50
	G-C (EG)	-7.90	-8.13	-8.84	-8.00	-2.90
	A-C (EG)	-6.70	-8.33	-9.54	-2.43	-0.51
	T-G (EG)	-6.20	-7.81	-9.87	-3.52	-3.91
	C-G (EG)	-7.70	-8.54	-9.03	-6.51	-8.80
	A-G (EG)	-6.50	-9.33	-8.80	-6.58	-4.49
	C-G (EG)	-12.40	-12.71	-13.11	-18.06	-10.44
	G-C (EG)	-11.60	-12.74	-12.96	-10.84	-16.40
	MAE	–	1.77	2.09	6.12	7.37

*The geometries of both GC WC (EG) pairs are identical.

Table S2: Interaction energies (kcal/mol) for the stacked base pairs with the intercalated phen ligand. A-T/phen/T-A MG and A-T/phen/T-A mg corresponds to Adenine-Thymine base pair system with intercalated phen in the major groove (MG) and minor groove (mg), while G-C/phen/C-G MG and G-C/phen/C-G mg corresponds to Guanine-Cytosine base pair system with phen intercalated in the major groove (MG) and minor groove (mg).

System	DLPNO-CCSD(T)	LMKLL	PM6-DH2	PM7
A-T/phen/T-A MG	-37.53	-38.26	-7.58	-3.97
A-T/phen/T-A mg	-33.81	-36.39	-5.71	-1.70
G-C/phen/C-G MG	-42.06	-41.69	-11.91	-8.03
G-C/phen/C-G mg	-35.87	-35.80	-6.49	-2.31
MAE		0.94	29.40	33.32

Table S3: Interaction energies (kcal/mol) for different DNA base pairs. Abbreviations used in the first column: A, T, C, G – adenine, thymine, cytosine, guanine; m – methyl-; WC, H – Watson-Crick, Hoogsteen: OG, EG – optimized geometry, experimental geometry. The third column correspond to the DFT calculations with the optimized pseudopotential and basis set. For the fourth and fifth column psml pseudopotential have been used but, in the fourth column optimized basis sets were used and for the fifth the default basis sets were used. The Mean Absolute Error (MAE) was calculated taking the experimental data as reference values.

| | System | Reference | LMKLL | LMKLL/psml | |
				Opt. Basis	Def. Basis
H-bonded base pairs	G-C WC (OG)	-32.06	-32.18	-32.18	-33.77
	mG-mC WC (OG)	-31.59	-32.09	-32.21	-33.75
	A-T WC (OG)	-16.86	-17.97	-18.16	-19.13
	mA-mT WC (OG)	-18.16	-18.84	-19.08	-20.56
	A-T WC (OEG)	-16.40	-17.95	-18.21	-19.06
	G-C WC * (EG)	-35.80	-34.84	-35.05	-36.62
	A-T WC (EG)	-18.40	-19.93	-20.13	-20.83
	G-A HB (EG)	-11.30	-13.67	-14.62	-14.99
	C-G WC (EG)	-30.70	-32.11	-32.21	-34.38
	G-C WC (EG)	-31.40	-32.02	-32.05	-34.35
	MAE	–	1.08	1.27	2.48
Stacked base pairs	G-C (OG)	-19.02	-21.03	-21.20	-23.14
	mG-mC (OG)	-20.35	-21.72	-22.04	-23.74
	A-T (OG)	-12.30	-14.25	-14.70	-16.01
	mA-mT (OG)	-14.57	-17.37	-17.79	-19.04
	A-T (EG)	-8.10	-11.89	-12.18	-13.46
	G-C (EG)	-7.90	-8.84	-8.83	-11.46
	A-C (EG)	-6.70	-9.54	-9.90	-11.05
	T-G (EG)	-6.20	-9.87	-10.26	-11.24
	C-G (EG)	-7.70	-9.03	-9.31	-10.49
	A-G (EG)	-6.50	-8.80	-9.36	-10.44
	C-G (EG)	-12.40	-13.11	-13.30	-15.16
	G-C (EG)	-11.60	-12.96	-13.25	-14.70
	MAE	–	2.09	2.40	3.88

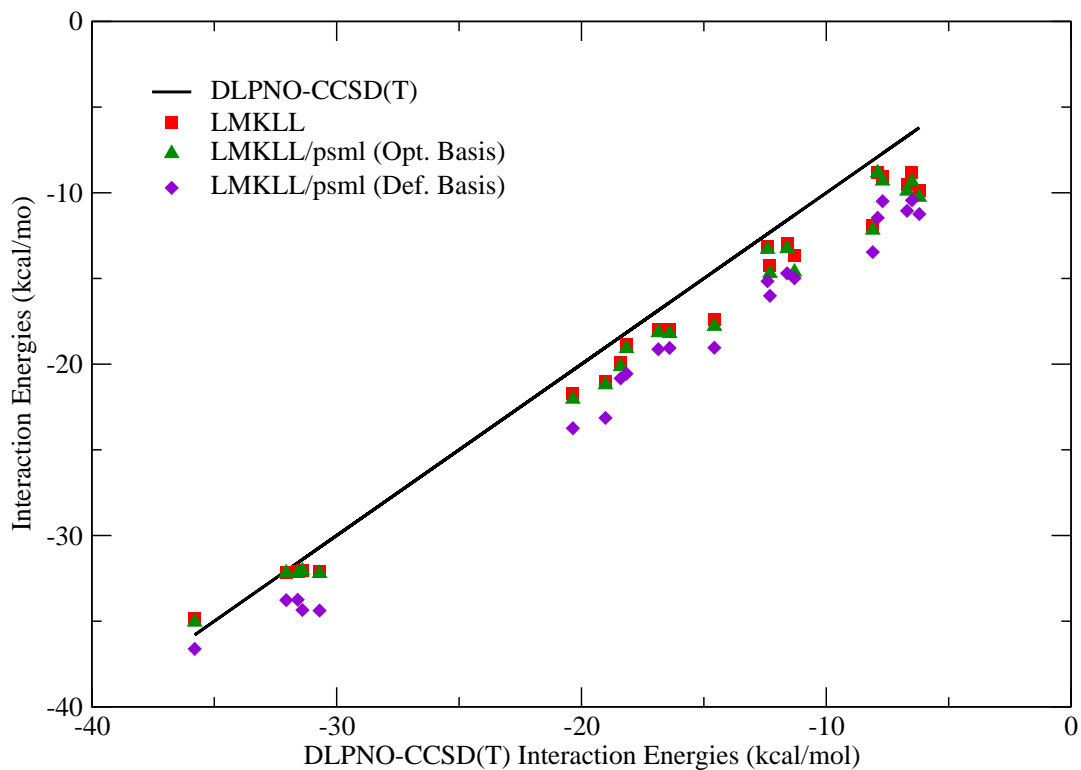*The geometries of both GC WC (EG) pairs are identical.

Figure S4: Interaction energies (kcal/mol) for the DNA base pair benchmark data set structures[1] with psml pseudopotentials. The LMKLL label correspond to the DFT calculations with the optimized pseudopotential and basis set. In the case of LMKLL/psml (Def. Basis) the psml pseudopotential were used with the optimized basis sets and, in the case of LMKLL/psml (Def. Basis) default basis were used. r^2 value for the LMKLL, LMKLL/psml (Def. Basis), and LMKLL/psml (Def. Basis) is 0.997, 0.996, and 0.995, respectively.

Table S4: Interaction energies (kcal/mol) of the stacked base pairs with the intercalated phen ligand. A-T/phen/T-A MG and A-T/phen/T-A mg corresponds to Adenine-Thymine base pair system with intercalated phen in the Major groove (MG) and minor groove (mg), while G-C/phen/C-G MG and G-C/phen/C-G mg corresponds to Guanine-Cytosine base pair system with phen intercalated in the Major groove (MG) and minor groove (mg). The third column corresponds to the LS-DFT calculations with the optimized pseudopotential and basis set. For the fourth and fifth column psml pseudopotential have been used but, in the fourth column optimized basis sets were used and for the fifth the default basis were used. The Mean Absolute Error (MAE) was calculated taking the DLPNO-CCSD(T) energies as reference values.

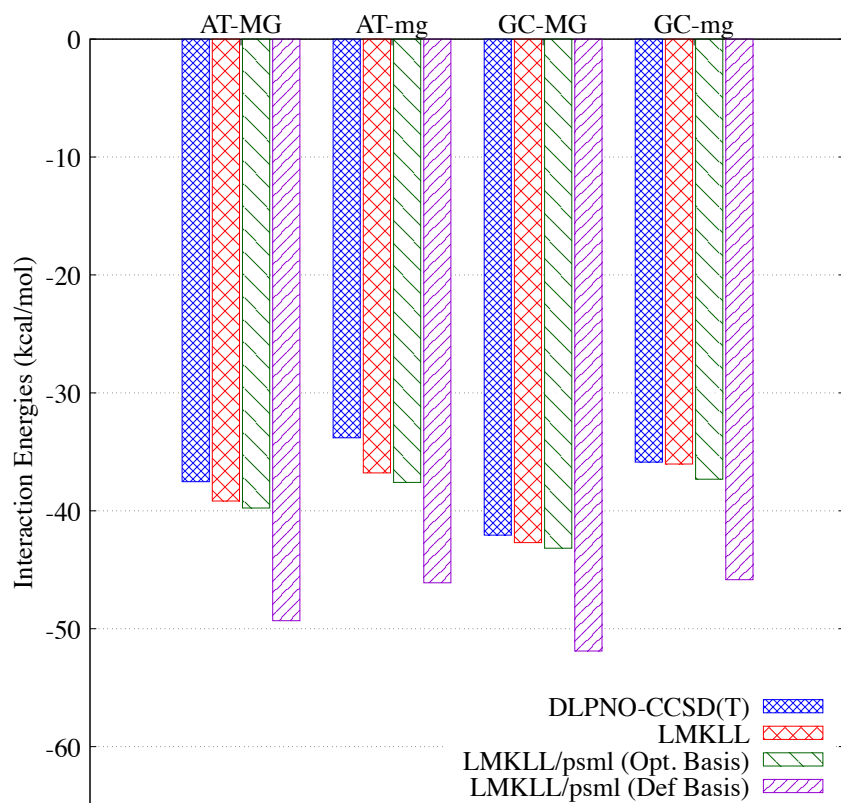| System	Reference	LMKLL	LMKLL/psml	
			Opt. Basis	Def. Basis
A-T/phen/T-A MG	-37.53	-39.17	-39.76	-49.32
A-T/phen/T-A mg	-33.81	-36.79	-37.59	-46.11
G-C/phen/C-G MG	-42.06	-42.70	-43.20	-51.89
G-C/phen/C-G mg	-35.87	-36.03	-37.33	-45.84
MAE		0.94	2.15	10.97

Figure S5: Interaction energies (kcal/mol) for the phen/DNA system with psml pseudopotentials. The LMKLL label corresponds to the LS-DFT calculations with the optimized pseudopotential and basis set. In the case of LMKLL/psml (Opt. Basis) the psml pseudopotential were used with the optimized basis sets and, in the case of LMKLL/psml (Def. Basis) default basis were used.

Table S5: Interaction energies (kcal/mol) for the different G-quadruplex structures (G_4MG_4, aG_4MG_4, GQM, and $GQ_{4Na}M$) with the different metal cations (Li, Na, K, Rb, and Cs) for the used computational methods (PM6-DH2, PM7, LMKLL/DZDP, and DLPNO-CCSD(T)/def2-SVP) along with the results found in the bibliography[2] at ZORA-BLYP-D3(BJ)/TZ2P level. The Mean Absolute Error (MAE) was calculated taking the DLPNO-CCSD(T)/def2-SVP energies as reference values.

	System	ZORA-BLYP-D3(BJ)	DLPNO-CCSD(T)	LMKLL	PM6-DH2	PM7
	Li	-161.50	-153.67	–	-101.62	-117.22
	Na	-152.10	-149.87	-134.56	-127.28	-122.41
G_4MG_4	K	-128.80	-129.86	-119.03	-78.80	-105.82
	Rb	-115.50	-115.54	-108.90	-67.96	-113.17
	Cs	-99.60	-97.20	-93.62	-97.89	-79.50
	Na	-145.80	-143.13	-129.79	-123.10	-118.35
aG_4MG_4	K	-126.60	-129.38	-118.00	-73.78	-102.97
	Rb	-114.70	-116.73	-108.84	-65.95	-102.83
	Cs	-99.20	-95.34	-93.30	-93.43	-78.05
	Li	-165.70	-158.36	–	-102.40	-110.31
	Na	-156.60	-153.93	-137.09	-113.40	-113.72
GQM	K	-134.70	-134.41	-123.74	-67.00	-107.12
	Rb	-119.10	-115.95	-115.17	-64.15	-117.80
	Cs	-104.40	-102.58	-101.09	-99.72	-94.30
	Na	-170.90	-170.04	-152.29	-115.60	-132.87
$GQ_{4Na}M$	K	-148.80	-148.73	-138.66	-88.68	-132.47
	Rb	-137.30	-136.45	-129.67	-77.25	-123.23
MAE		2.47	–	9.04	40.86	22.51

Table S6: Total Energy (eV), Wall Time (s) and RMSD (Å) for the G_4MG_4 system[2] geometry optimization with different max force tolerance. RMSD value was calculated taking as reference structure the geometry of the literature.

Max Force Tolerance	Total Energy	Wall Time	RMSD
0.5	-4659.90	19171.3	0.01
0.2	-4659.97	19287.7	0.02
0.1	-4659.99	21431.6	0.02
0.07	-4660.00	23733.8	0.03
0.05	-4660.00	31654.6	0.06
0.02	-4660.01	40450.2	0.07

References

1.- Jurečka, P.; Šponer, J.; Čern'y, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA

base pairs, and amino acid pairs. Physical Chemistry Chemical Physics 2006, 8, 1985–1993

2.- Zaccaria, F.; Paragi, G.; Guerra, C. F. The role of alkali metal cations in the stabilization of guanine quadruplexes: why K+ is the best. Physical Chemistry Chemical Physics 2016, 18, 20895–20904