# 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[a][b], Sawssen Elleuchi[c], Khaled Jarraya[c], Emilio Artacho[a][d][e][f], Xabier Lopez[b][f]*, and Adrià Gil[a][g][h][i]*

[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*

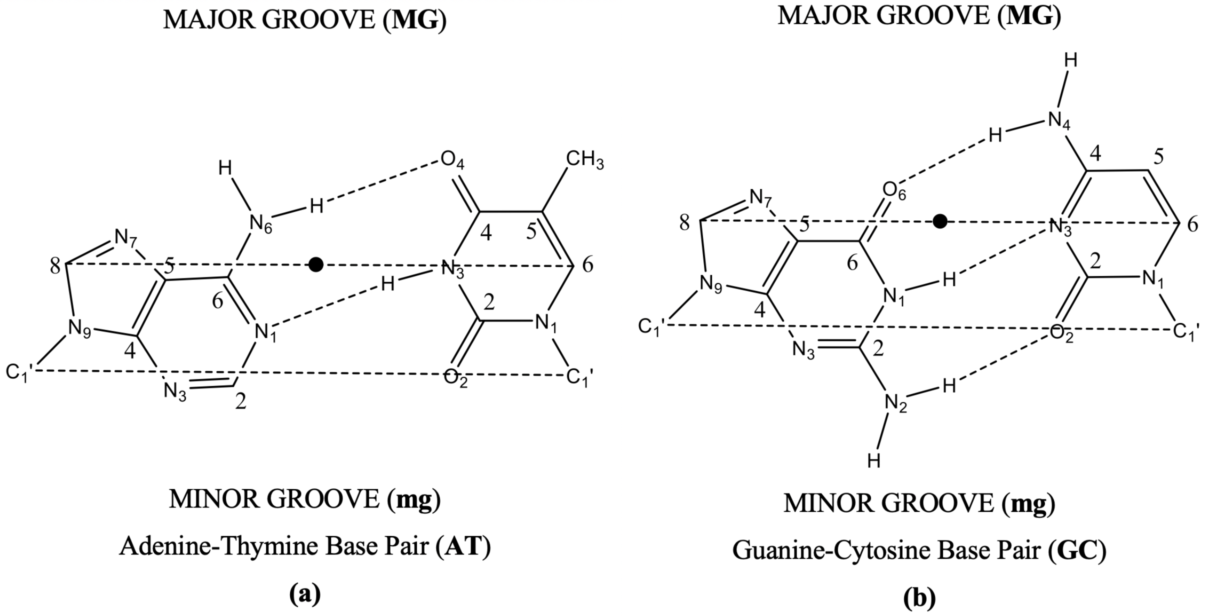
[h] *Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH) CSIC- Universitad de Zaragoza, C/ Pedro Cerbuna 12, 50009, Zaragoza, Spain.*

[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 ($\vartheta$) 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 $\vartheta$ 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 $\vartheta$ 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 $\vartheta$ 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.



Adenine-Thymine Base Pair (**AT**)

**(a)**

Guanine-Cytosine Base Pair (**GC**)

**(b)**

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 ($\vartheta$) 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 ($\vartheta$) 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 $\vartheta$ 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.

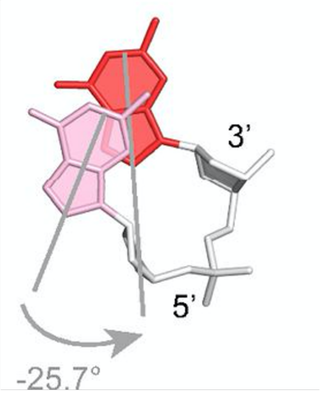


Figure S2: Representation of the stacking of two G-tetrad guanines.The $\vartheta$ twist angle, is defined as the angle between the lines formed by the guanine $C_8$ and the midpoint between $N_2$ and $C_6$.

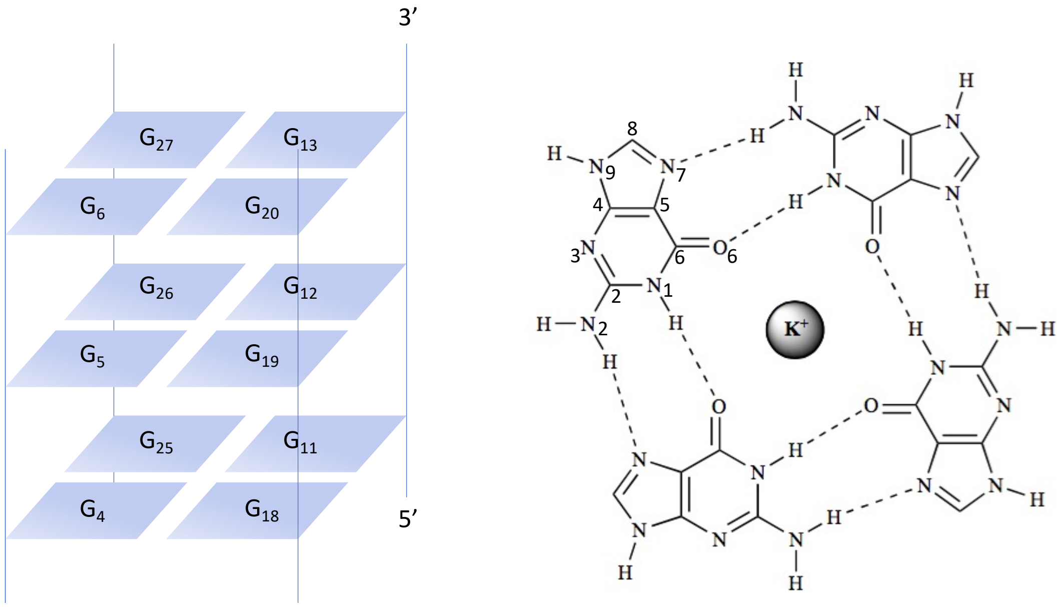


Figure S3: On the left, scheme of the stacking of three G-tetrads of guanines composing the 2JWQ PDB structure, and nomenclature for each guanine base. On the right, a G-tetrad, with the number used in the nomenclature.

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, HB – 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, HB – 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 (EG) | -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.

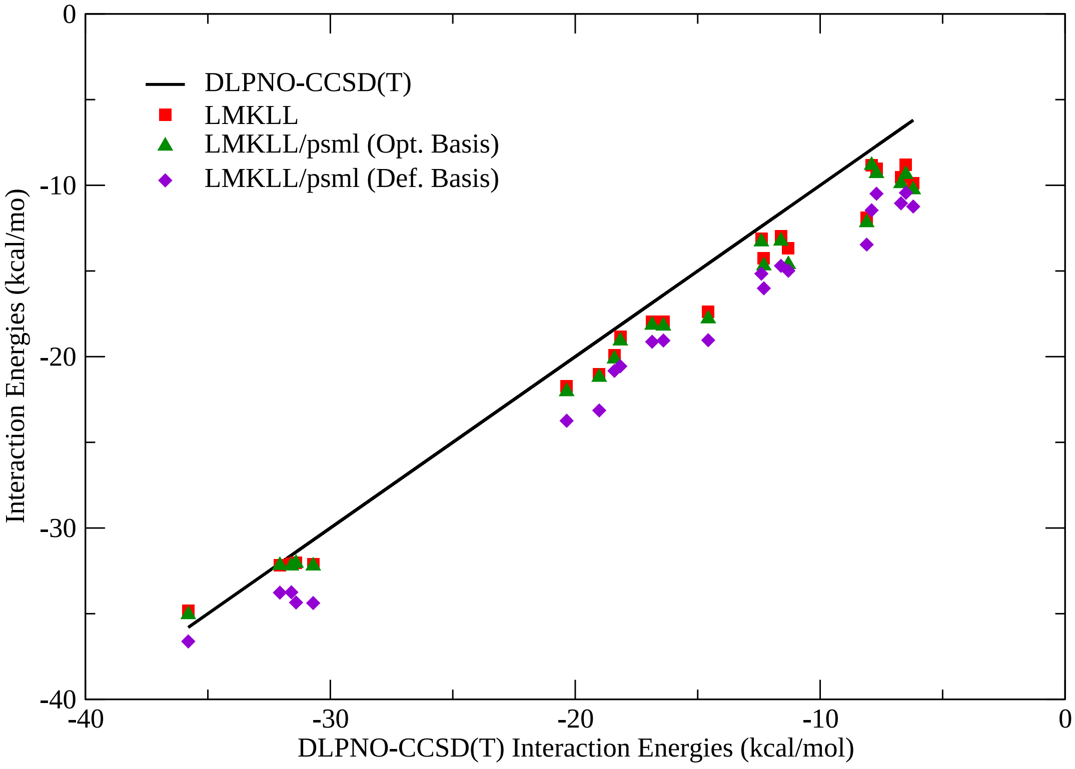


Figure S4: Interaction energies (kcal/mol) for the DNA base pair benchmark data set structures[1] with psml pseudopotentials. The LMKLL label correspond to the DFT calculations with the optimized pseudopotential and basis set. In the case of LMKLL/psml (Def. Basis) the psml pseudopotential were used 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 |

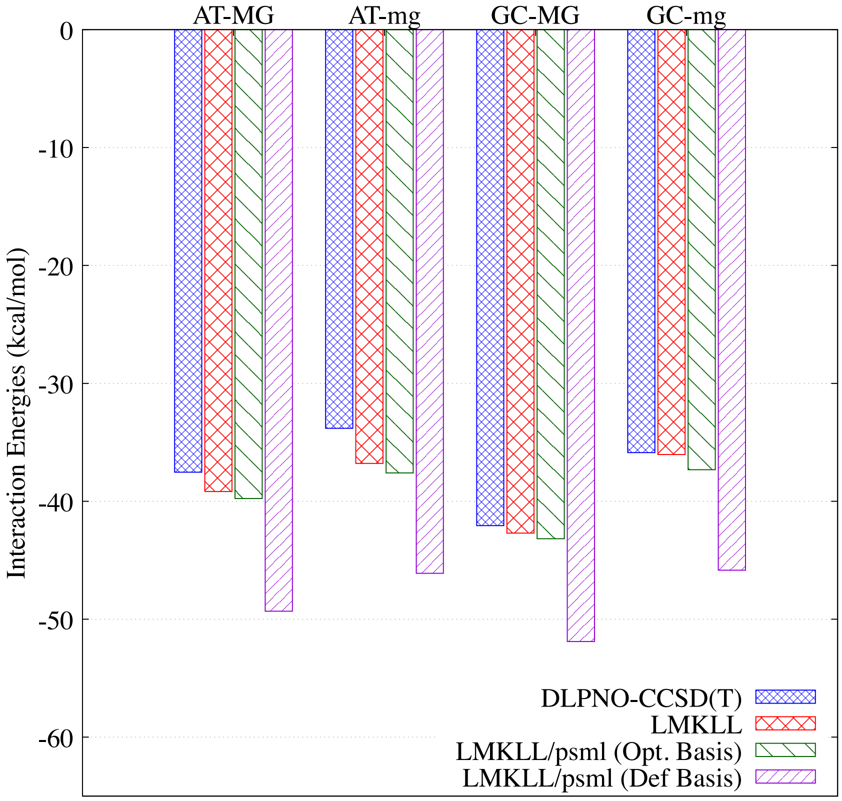


Figure S5: Interaction energies (kcal/mol) for the phen/DNA system with psml pseudopotentials. The LMKLL label corresponds to the LS-DFT calculations with the optimized pseudopotential and basis set. In the case of LMKLL/psml (Opt. Basis) the psml pseudopotential were used with the optimized basis sets 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 |
|---|---|---|---|---|---|---|
| $G_4MG_4$ | Li | -161.50 | -153.67 | – | -101.62 | -117.22 |
| | Na | -152.10 | -149.87 | -134.56 | -127.28 | -122.41 |
| | 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 |
| $aG_4MG_4$ | Na | -145.80 | -143.13 | -129.79 | -123.10 | -118.35 |
| | 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 |
| GQM | Li | -165.70 | -158.36 | – | -102.40 | -110.31 |
| | Na | -156.60 | -153.93 | -137.09 | -113.40 | -113.72 |
| | 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 |
| $GQ_{4Na}M$ | Na | -170.90 | -170.04 | -152.29 | -115.60 | -132.87 |
| | 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 |

Table S7: Computation Wall-time (seconds) for the 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). DLPNO-CCSD(T) and LMKLL calculations were computed with 24 processors belonging too a single node. The PM6-DH2 and PM7 calculations were performed in a single processor. The used processors were Intel Xeon E5-2683 v4.

| | System | DLPNO-CCSD(T) | LMKLL | PM6-DH2 | PM7 |
|---|---|---|---|---|---|
| $G_4MG_4$ | Li | 2352.7 | – | 0.6 | 1.4 |
| | Na | 2061.7 | 6872.8 | 0.7 | 0.6 |
| | K | 4227.7 | 12064.4 | 0.6 | 0.8 |
| | Rb | 2887.5 | 5568.7 | 0.8 | 0.7 |
| | Cs | 2546.0 | 8261.4 | 0.8 | 0.8 |
| $aG_4MG_4$ | Na | 2474.3 | 13696.2 | 0.9 | 0.9 |
| | K | 2284.8 | 9183.3 | 0.9 | 0.9 |
| | Rb | 1941.3 | 6849.5 | 0.9 | 0.9 |
| | Cs | 1659.6 | 4078.8 | 0.9 | 0.9 |
| GQM | Li | 4531.2 | – | 5.0 | 11.5 |
| | Na | 4936.9 | 8150.6 | 8.2 | 6.3 |
| | K | 5121.8 | 13695.2 | 3.5 | 10.1 |
| | Rb | 7422.2 | 8286.0 | 3.1 | 5.1 |
| | Cs | 7385.4 | 6535.8 | 3.3 | 4.8 |
| $GQ_{4Na}M$ | Na | 8005.6 | 3849.9 | 3.3 | 7.0 |
| | K | 8243.4 | 5485.1 | 3.3 | 6.0 |
| | Rb | 9014.7 | 5480.0 | 3.0 | 6.2 |

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