AMOEBA Polarizable Molecular Dynamics Simulations of Guanine Quadruplexes: from the c-Kit Proto-oncogene to HIV-1

Dina S. El Ahdab,[†,‡,¶] Louis Lagardère,[†] Zeina Hobaika,[†,‡] Théo Jaffrelot Inizan,[†,§] Frédéric Célerse,[†] Nohad Gresh,[†] Richard G. Maroun,[†,‡] and Jean-Philip Piquemal[*,†]

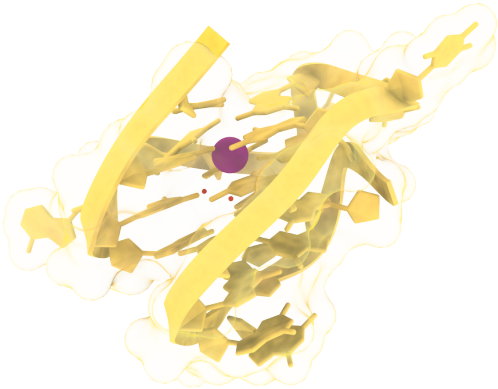
†*Sorbonne Université, Laboratoire de Chimie Théorique, UMR 7616 CNRS, 75005, Paris, France*

‡*Equipe Structure et Interactions des macromolécules, UR EGP, Centre d'Analyses et de Recherche, Faculté des Sciences, Université Saint-Joseph de Beyrouth, Beirut, Lebanon*

¶*Qubit Pharmaceuticals, 75014 Paris, France*

§*Present adress: University of California Berkeley, Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, Berkeley 94720, USA*

E-mail: jean-philip.piquemal@sorbonne-universite.fr



Abstract

Long oligomer sequences, rich in guanine and cytosine, such as *c-kit1* and the HIV-1 LTR-III sequence, are prevalent in oncogenes and retroviruses and play crucial roles in cancer. Understanding the conformational dynamics of such guanine quadruplexes and identifying druggable regions are therefore essential for developing new inhibition strategies. In this study, we used extensive AMOEBA polarizable force field molecular dynamics simulations combined with data-driven adaptive sampling and clustering algorithms, reaching a cumulative simulation time of 7.5 μs for *c-kit1*. Our results reveal novel structural stabilizations and flexible loop dynamics, as well as the critical role of polarizable water in stabilizing the G-quadruplex, leading to the identification of two new druggable pockets in *c-kit1*. Additionally, 400 ns simulation of the HIV-1 LTR-III sequence confirmed its quadruplex stability and uncovered a potentially druggable cryptic pocket.

Introduction

Long oligomer sequences rich in Guanine-Cytosine (GC) are abundantly represented in oncogenes and retroviruses.[1,2] Their propensity to reorganize from double to quadruple helices was long recognized.[3,4] This occurs after strand separation and, in the G-rich strand, the formation of stacked cyclic guanine tetramers involving well-defined G bases that can be distant in the primary sequence.[5] Such cyclic G tetramers were identified early on[6] and are

stabilized by Hoogsteen pairing: each G has its O6 and N7 atoms accepting a proton from N1 and N2 of a neighboring G and in turn donates a proton from its N1 and N2 atoms to O6 and N7 of its other G neighbor.[6] A recurring motive in Guanine-Quadruplexes (GQ) is that of three stacked tetrads connected together by one or more nucleotide loops, as represented in figure 1. Alkali metal cations, mostly K^+ and Na^+, can be essential contributors to GQ stability.[7–10] There are two cations located along the z axis perpendicular to the three successive stacked tetramers, close to their geometrical centers.[11,12] GQs were recognized early on as targets for the design of chemotherapeutic ligands.[2] This is the case for human telomers which in cancer cells do not undergo shortening after each cell division, thus improving the immortalization of these cells.[1] Such a shortening is prevented by the telomerase enzyme.[13] Thus its inhibition by high-affinity GQ-binding ligands could oppose immortalization.[1] GQ's are involved in the regulation of gene expression at the transcriptional and translational levels.[1] They are also found in the gene promoters of a wide range of genes that are essential in cell signalling, which modulates the six hallmarks of cancer.[14–17] We report in this work multi-μs polarizable molecular dynamic simulations bearing on two GQs, a human and a retroviral one, to unravel their distinct structural features. The first GQ is c-Kit, a proto-oncogene encoding for the stem cell factor receptor CD117. CD117 is a tyrosine kinase receptor involved in intracellular signaling with consequences in differentiation, survival, adhesion, motility, angiogenesis and proliferation.[18] When mutated, CD117 can be responsible for several cancers, such as gastrointestinal stromal tumor (GIST), acute myeloid leukemia, and melanoma.[19] While an imatinib mesylate tyrosine kinase (TK) inhibitor has been found to be quite effective on GIST patients, it didn't achieve an optimal response due to inhibitor resistance.[18,20] Thus, interference with TK transcription at the genomic level could be attempted instead, such as by early stage inhibition of *c-kit1* transcription into CD117 mRNA.

Augmenting the stability of *c-kit1* proto-oncogene GQ by high-affinity ligands, by competing with its binding to the polymerase, could therefore constitute a novel transcription

therapy against cancer. We consider the *c-kit87up, 22-mer* sequence; d(AGGGAGGGCGCTGGGAGGAGGG), which is 87 nucleotide upstream of the transcription start site of the c-kit gene in human genome. It forms a single GQ species in K^+ containing solution.[21] The presence of four tracts of three consecutive guanines separated by one up to four-residue linkers results into folding in a conventional parallel-stranded GQ structure. As shown in Figure 1, the first tetrad is formed by nucleotides G2, G6, G10 and G13. The second tetrad is formed by G3, G7, G14 and G21. The third is formed by G8, G14, G15 and G22. NMR and circular dichroism experiments showed the connecting linker sequences to be essential for GQ core stabilization.[21] These linkers are formed by five loops: two single-nucleotide loops with A5 and C9; one two-nucleotide lateral loop with C11 and T12 and one five-nucleotide long propeller-loop (LP) with A16, G17, G18, A19 and G20. High-resolution structures of *c-kit1* GQ in the presence of K^+ are available from 3D-NMR spectroscopy (PDB code 2O3M) and X-ray crystallography (PDB codes: 3QXR, 4WO2, and 4WO3).[11,22,23] Since 4WO2 has the best crystallographic resolution for the 22-mer structure (1.82 Å) of human native c-Kit proto-oncogene promoter, it was selected as a starting structure for this study.

This study will be then extended to G quadruplexes of the HIV-1 Long-Terminal Repeat 3 (LTR-III). While *c-kit1* lent itself to in-depth MD analyses in the context of classical potentials[12] and Drude polarizable potentials,[24] there are no precedents to our knowledge to MD simulations on any HIV-1 forming G-quadruplex. This extension will allow to unravel the similarities and differences between c-kit-1 and HIV-1 quadruplexes. Several G-rich sequences in the U3 promoter region of the HIV-1 Long Terminal Repeat (LTR) are shown to fold into a range of dynamically interchangeable G-quadruplex structures.[25–27] These sequences regulate viral transcription through the stabilization of LTR GQ by GQ-specific ligands. Among all GQs formed in the LTR sequence, the specific LTR-III was shown to be the GQ giving rise to the richest conformational diversity *in vitro*. Thus, an NMR structure of LTR-III in K^+ solution, unraveled the following motives: i) a unique quadruplex–duplex

hybrid consisting of a three-layer (3 + 1) GQ scaffold; ii) a 12-nt diagonal loop containing a conserved duplex-stem where six nucleotides are interacting by W-C hydrogen bonds to form a hairpin; a 3-nt lateral loop and a 1-nt propeller loop[24] (Figure 1).

Crystal and NMR structures represent the most statistically favorable conformations and do not provide information about conformational dynamics, which is often insufficient for drug design, thus necessitating the use of large-scale molecular dynamics simulations. To do so, we will use the extension AMOEBA polarizable potential to nucleic acids,[28] enabling to reliably translate the stability conferred by the presence of two K^+ cations bound in the center of the two successive stacked G tetramers. Indeed, the use of polarizable force field has been shown for be particularly suited for complex and charged biological environments.[29–35] Worth mentioning in this connection is a recent validation study with the SIBFA polarizable potential[36]demonstrating that it is possible to closely reproduce the Quantum Mechanical (QM) energy profile for the channeling of alkali cations along the z axis of two stacked G tetramers[37] while this was not possible with non-polarizable force fields.[38]

We leveraged the latest AMOEBA force field through the GPU-accelerated Tinker-HP[39,40] software and the use of unsupervised adaptive sampling. Such a strategy[32] has been shown to greatly enhance the conformational sampling of the SARS–CoV–2 Main Protease[32,33] and Spike[35] proteins, offering access to multi-μs MD simulations. For GQs, this approach is particularly suited for an in-depth understanding of the c-kit G-quadruplex that requires particulary long simulations. Indeed, overstabilizing the quadruplex conformation of the *c-kit1* proto-oncogene could impair, at an early genomic stage, RNA polymerase-mediated initiation of transcription.[1] This could enable to circumvent resistance to tyrosine kinase inhibitors which takes place at the later proteomic stage. Our *c-kit1* simulations will be compared with recent results obtained from AMBER[12] and CHARMM C36.[24] Regarding *c-kit1*, these comparisons now unveil previously unnoticed structural features. These include differences in clustering, the evolution of loops, responses to cations in the ionic channel and the emergence of alternative transient states.

Results and discussion

We will first study the *c-kit1* conformational stability and transition states including the loop conformation then we will focus on clusters analysis. We will propose a detailled study of the GQ stabilization by K^+ and polarizable H_2O molecules. After extending the simulations to another related GQ present in HIV-1, denoted LTR-III, we will conclude by unravelling potential cryptic pockets in both c-kit1 and LTR-III structures..

Conformational stability and transition states

Evolution of the stability of the different GQ *c-kit1* nucleotides

To study the evolution of the stability of the different GQ *c-kit1* nucleotides, the root mean square deviation (RMSD) was calculated for all of the 7.5 μs of adaptive sampling trajectories in order to measure the average distance between the backbone atoms in the superimposed structures. It amounts to 3.5 Å for the whole GQ structure Figure 2 a), indicative of its overall stability. However, in order to follow the fluctuation dynamics of the different GQ components, the RMSDs for the GQ core, the LP loop and for the single nucleotide loops were calculated separately. Both the GQ core and the LP loop having RMSDs below 1 Å appear stable and structured. On the other hand, the single nucleotide loops have RMSDs fluctuating between 4 and 7 Å thus accountable for the whole structural dynamics.

The root mean square fluctuations (RMSF) were also calculated to search for the most dynamic *c-kit1* GQ bases. These were found to be single nucleotides A1, A5, C9, and dinucleotide C11-T12 (figure 2, b). It is thus the flexibility of these bases that most impacts the entire GQ dynamics. A better understanding of the dynamics of these nucleotides is enabled by the corresponding temporal evolutions of two torsion angles, four distances and of the *pi–pi* stacking indices as reported below.

Evolution of the loop conformation

We then investigated the evolution of the loop conformation. To do do, we have monitored the χ glycosidic torsion angle evolutions, with the possible occurrence of *(anti / syn)* base flipping. The second monitored torsional angle is the angle γ governing the orientation of the sugar with respect to the phosphate backbone. Figure 1 and Figure 2 in the SI represent a 2D plot of χ and γ angle fluctuations of A1, A5, C9, C11 and T12. C9 and C11 display large amplitudes of χ and γ variations compared to A1, A5 and T12. A1 and T12 have limited χ angle variations (-50° – +50° and +25° – +50° respectively) but have large γ angle fluctuations. Such fluctuations enable the formation of a transient Hoogsteeen base-pair above the first GQ tetramer involving G2, G6, G10, and G13. It is stabilized by two H-bonds, between HN6(A1) and O4(T12), and between N7(A1) and HN3(T12). The involvement of N7 of A1 instead of N1 as in a standard W-C base-pair was unanticipated and appears to have no precedents from other GQ simulations. It locates N7 of A1 in the same groove as the three G bases, G2, G3 and G4 on the same strand which form the three tetramers. By contrast, it is the formation of a W-C base-pair that was previously reported in the AMBER MD simulations. This CMD simulation gave rise to the formation of a competing, longer-duration base-pair between A1 and C11. As an estimate, the time evolutions of these two base-pairs were monitored. We considered first the initial CMD trajectory (800 ns), then trajectories resulting from adaptive sampling (7.5 μs). Figure 3 reports the CMD evolutions of the HN6–O4 distance for A1–T12 and of the HN6–N3 distance for A1–C11. As shown in Figure 3, the first shortening of the first distance to app. 2 Å is enabled by the flip movement of T12 with large amplitude γ fluctuations, and its stacking above the open face tetrad, particulary above G13 and G10. It first occurs after 80 ns but only lasts 1 ns. It occurs again between 180 and 220 ns. At 260 ns, the formation of the concurrent A1-C11 base-pair takes place and this base-pair lasts till the end of the CMD at 800ns. Such a non-classical interaction was not previously reported in polarizable force field nor in non-polarizable force field studies simulations.[12,24] However, AMBER MD[12]

did record a movement of C11 over the helix terminal along with a short-lived interaction with either T12 or the phosphate backbone of G10. C9 and A5 remain highly dynamic and do not interact with any particular GQ nucleotide because of their locations in single nucleotide lateral loops. Within the stable LP loop two stable interactions can be detected, namely between A16 and G20, and between G17 and A19. We have monitored the time evolutions of one relevant intermolecular distance, between A16(N1) and G20(HN1) and between G17(HN2) and A19(N1). Such distances undergo only limited variations around 2 Å for A16–G20 and 2.15 Å for G17–A19, attesting to a lasting stability of the LP loop.Such distances were recorded to vary between 2 and 3 Å respectively in AMBER MD.[12] This result suggests that AMOEBA is able to detect and maintain non-classical interactions between nucleic bases. Their onset starts after 10ns. In contrast, an A1 flip movement was detected only after 2 μs in a MD simulation in the AMBER simulations.[12]

A1 is found to interact alternatively with G2 and G13, while C11 and T12 interact with G6 and G10. Since the behavior of highly dynamic loops shows flip movements from the X-ray conformation and π-stacking interactions with the open face tetrad, we calculated the stacking index of the four bases in order to unravel the bases involved in the stacking. G2 is thus found to interact with A1 while both C11 and T12 interact with G10, Figure 3. Thus the stability of A1 above the open face tetrad could be favored by two types of interaction : a short-lived Hoogsten base-pair (*cis* with T12 and a long-lived *anti* WC base-pair with C11 together with a π-stacking interaction with G13 .

tICA clustering and cluster analyses

tICA is a dimensionality-reducing procedure[41–44] whose goal is to identify the slowest degrees of freedom (with the largest autocorrelation) of trajectories. It enables to identify microstates characteristic of the MD trajectories which result from the unsupervised adaptive sampling strategy. Such microstates are then regrouped into macro-states representative of each cluster. As exposed in our previous work,[32] DBSCAN[45] is specially designed to iden-

tify clusters of arbitrary shape. Using two parameters: ϵ, the distance at which two points are considered neighbors, and *MinPts*, the minimum number of points necessary to define a cluster. DBSCAN unravels the existence of seven independent clusters as shown in Figure 4. The interest of clustering lies in the aim of being able to understand the various structural changes that may exist within each group of structures, thus allowing a better characterization of conformations in order to subsequently facilitate the design of future stabilizing molecules.

Thus, we computed the relative free energy with respect to the most representative cluster (cluster 1). Such energies are computed through the evaluation of the probability distribution over the total number of structures. According to Figure 5 and Table 1 in the SI, cluster 1 has the lowest free energy thus the highest probability of existence. Cluster 5 ranks second with less than 1 kcal/mol higher free energy, the other clusters having 1.5-2.5 kcal/mol higher free energies. The RMSFs of the seven clusters enable to quantify their dynamics within the GQ. Figure 3 (SI) shows that nucleotide A1, single nucleotide loops A5 and C9 and dinucleotide loop C11-T12 undergo large fluctuations. Except for A5, these have unequal amplitudes in the clusters, in the range 6-9 ÅḂy constrast, the LP loop and the all twelve G bases of the GQ core undergo less than 2 Å variations. A1, A5, C9, C11 and T12 are thus identified by the RMSFs on the seven tICA-derived clusters as the most flexible nucleotides of *c-kit1* GQ, giving their individual shapes to the clusters. It is instructive to identify some of their key evolving structural features in these clusters. We have thus first considered the two above-mentioned torsion angles,χ and γ. Regarding χ angles, Figure 1 in the SI shows that all clusters share the same highest probability densities for A1, A5, C9 and T12. For C11, the χ fluctuations in cluster 1 are centered around a Gaussian function peaking at $\tilde{1}0°$. In the other clusters its χ fluctuations are centered around two major states with peaks in two zones: -120°—170° and 120°–170°. Regarding γ angles, Figure 3 in the SI shows that all clusters have their highest probability densities for A1 and A5 centered around -25° and -110 ° respectively. Major differences between clusters are seen for C9, C11

and T12. The highest probability density for the C9 γ angle is located between 75° and 130° for clusters 1, 3, 6 and 7. Clusters 2, 4 and 5 have their peak at the same -40° value for this base. C11 is highly structured in cluster 6 but in cluster 1 its γ angle has two peaks, around -25° and 80°. In the other clusters, the γ probability density has shallow variations in two zones, the first between -100°—180° and the second between 100°–180°. For T12, γ peak around ~50° for clusters 1, 2, 4, 5 and 7 and around 60° for clusters 3 and 6. These results show that C11 can undergo continuous changes of its χ angle in the different clusters enabling it to populate both *syn* and *anti* states. By contrast, A1, A5 , C9 and T12 have χ centered around a single peak in all clusters. Regarding γ, both A1 and A5 show values centered around one single peak in all clusters. Both C9 and T12 show two predominant values, centered around -40° and 75°. Finally C11 has one peak in cluster 6, and two peaks in all other clusters. These observations demonstrate the high stability of the GQ core and the LP loop in *c-kit1* GQ structure regardless of the defined cluster. However it is clear that the single nucleotides loops (A5 and C9), the free nucleotide (A1) and the 2 nucleotides loop (C11 and T12) are highly flexible since they are not engaged in stable intra-molecular interactions.

Yet, both conventional and adaptative MD have identified transient H-bonding interactions between N1 of A1 and HN3 of C11. H-bond interactions were also detected between N1 of A1 and HN3 of T12 , but it is not maintained through the full time of MD simulations. The involved H-bonding distances were accordingly computed in the regrouped seven clusters. Figure 4 in the SI shows clearly that the highest probability density is reserved for a defined H-bond interaction betwen A1 and C11, whereas the interaction betwene A1 and T12 is not perfectly defined as for A1–C11. Additionaly, we have recently resorted to the π-π stacking index[46] in an analysis of the M^{pro} SARS–CoV–2 protein. It has enabled us to quantify the extent of stacking between two residues, Phe140 and His163, a key interaction which stabilizes an oxyanion hole structure upon which the activity of each M^{pro} protomer depends.[32] Such an analysis is pertinent in the present study, as it can similarly evaluate

the stability of the stacking of A1, C11 and T12 above the first guanine tetrad. At the outcome of adaptive MD simulation, Figure 6, a), shows A1 to interact by π π-interaction with G13 in a persistent mode. This is shown by a high probability density for a stacking index of 0.2. Figures 6, b) and c) show that C11 and T12 stack above G10 preferentially than above G6. Thus on the one hand, the stacking index for C11–G10 of 0.15 has a much higher probability density than the stacking index of 0.12 of C11–G6. On the other hand, the stacking index for T12–G6 is 0.10, smaller than for T12–G10 (0.18). These results confirm the rearrangement of the nucleotides A1, C11 and T12 in an attempt to form a new plan of nucleotides interactions to enhance, furthermore, the stability of *c-kit1* GQ. We beleive that there is a transitional mode of interaction between A1 and C11 on one hand and between A1 and T12 on the other hand. According to our observation, the interaction between A1 and C11 was more robust and better defined in terms of H-bond interactions. Adding that A1 and C11 have a better π stacking index with respect to the guanines forming the first plan of guanines tetrad. Understating, the dynamic of these rearangement is very crucial for the upcomping steps of drug design, because they should be taken into consideration for the incoming ligands targeting *ckit-1* GQ.

GQ stabilization by K^+ and polarizable H_2O molecules

GQ interact with ions in an extremely dynamic fashion causing experimental detection difficulties.[47] Nevertheless, based on structural techniques,[48] it was shown that tightly bound to G-quadruplex DNA are actually "NMR visible" in solution. In that contexte, it was put forth that K^+ cations inside the channel exert an essential stabilizing role to the *c-kit1* promoter and, indeed, a few high-resolution structures in the presence of K^+ are available from 3D-NMR spectroscopy and X-ray crystallography. As discussed, we retained as a starting structure the best crystallographic resolution (PDB: 4WO2[22]) for the 22-mer structure (1.82 Å) of human native c-Kit proto-oncogene promoter. In order to study the *c-kit1* GQ behavior in the absence of these K^+ cations, we performed 20 independent simulations of

20ns each, using a similar setup as described in the Appendix, starting from an empty ionic channel where we displaced the structural K^+ coordinates outside of the GA structure and to join the bulk ions.

We assessed the stability of the GQ structure by calculating its RMSD in each trajectory. Figure 5 in the SI shows clearly that *ckit1* GQ preserved its stability and initial conformation during all 20 trajectories with some minor conformational changes of the unengaged free nucleotides (A1, C11 and T12). GQ core stability can be explained by the fact that the ionic channel was never seen empty by the end of any of these MD simulations. We counted the number of H_2O molecules and K^+ cations inside the ionic channel for every frame of these 20 MD simulations to understand the process of the ionic channel filling. To do so, we defined a virtual sphere centered on the mass center of the 12 Guanines contributing to the formation of the *ckit1* GQ core and with a radius of 4 Å (which is the best radius to cover all and only the ionic channel). As it is shown in Figure 7, and more in details in Figure 6 in the SI, several scenarions of the ionic channel filling can be possible.The highest frame density detected refers to the presence of 1 H_2O molecule and 1 K^+ cation inside the ionic channel. The second highest density corresponds to an empty channel where no H_2O molecules, nor cations are detected inside the GQ core. We recall that no channel is empty at the end of any of the 20 MD simulations and the density here corresponds to the frames from the beginning of the trajectories in which the ionic channel is empty. All *ckit1* GQ ionic channel start to be fulfilled after a while from the beginning of the MD simulations. The third highest density observed, corresponds to 2 H_2O molecules and 1 K^+ cation that can be present inside the ionic channel. To a slightly lower density, we can observe either the presence of 2 H_2O molecules, either 1 H_2O molecule only.

Once the channel is thus filled, all loops evolve in the same fashion as in the initial MD simulations with two K^+ cations present since their start with no significant relative structural changes. We calculated the average dipole moments of the H_2O molecules that have entered the channel. Their mean value : 2.72 D is slightly below the 2.78 D AMOEBA value

computed for liquid H_2O. This, albeit modest, decrease could possibly be due to a local decrease of the polarizing guanine-induced field inside the channel, due to the mutually-opposing fields exerted in each tetramer by the two pairs of two symmetry-related guanines. In this connection, it was recently observed[33] that a dense and a compact environment generates a global many-body depolarizing effect influencing the H_2O molecule induced dipoles. CHARMM simulations resulted into one K^+ from the solvent phase entering the channel after 20ns.[24] This entrance was from below tetrad 3 rather than, as in AMOEBA MD, from above tetrad 1. It reached tetrad 3 after transiting through the Long Propeller Loop, which as mentioned in,[24] could cause transient conformational distorsions in the LP loop and tetrad 3. Drude simulations,[24] starting from an empty channel, resulted into one K^+ transiently interacting with T12 and then entering the GQ core through the first tetrad, as found here in the present work. The additional advantage AMOEBA is bringing here, is its capacity to show that even if K^+ cations are not entering the ionic channel, GQ can still be perfectly maintained thanks to the entry of water molecules.

Cryptic pockets

We used the DoGSite Scorer software[49] to see if potential drug binding cavities could be detected in any of the seven identified groups. For all seven clusters, two pockets could be defined: P1, formed by bases A1, G2, G3, G13, G14, G15, A16, and P2, formed by bases G7, G8, C9, G10, C11, G21 as shown in Figure 8. To further study how some key pockets inter-base distances persist throughout adaptive MD, and how do their values compare in the seven clusters, we have considered three pairs inside P1: G3–G15, G2–G14 and A1–G13 and two pairs inside P2: G21–G8 and G10–G7. In the G–G pairs, the first G belongs to one tetrad and the second to the following tetrad. It is stacked below the guanine that donates two H-bonds to the first G. The intermolecular distances which we monitored are between N7 of the first G and HN1 of the second. In the A–G pair, the corresponding distance is between N7 of A1 and HN1 of G13 which is part of the first tetrad. This choice of G–G

distances should enable to evaluate if partial unstacking of the quadruplexes or distortion of the GQ H-bonding could have occurred during MD. The A1–G13 distance is seen to undergo large fluctuations in all clusters, notably in clusters 5 and 7. This is clearly caused by the large-amplitude motion of A1, whose RMSF exceeds 9 Å in these two clusters. However, Figure 7 in the SI shows that in all seven clusters, for each G–G pair, such distances remain closely distributed around sharp peaks in a 4 - 4.5 Å interval. This attests to the stability of the GQ core in all clusters. It should favor the binding of an incoming ligand to the groove formed by the bases belonging to P1 or to P2.

Simulation of the LTR-III region of HIV

HIV-I Long Terminal Repeat (LTR) is a crucial regulatory region that controls viral transcription and replication. Since this regions is also prone to GQ-formation, we have performed a preliminary AMOEBA simulation of 400 ns CMD using a high-resolution NMR structure as a starting point (PDB ID : 6h1K). DoGSite Scorer detected two cryptic pockets: P1, formed by bases G2, A4, G5, G6, G8, T9, G11, C12, C13, T14, G15 and G19 and P2 formed by bases G1, G2, G3, G15, G19, G20 and G21. P1 is mainly located in the hairpin domain while P2 is located in the upper face of GQ domain (Figure 9 a and b)). Similar to *c-kit1* GQ, RMSD and RMSF calculations (Figure 8,a and 8,b respectively in the SI) attest to a highly structured GQ core. By contrast, the hairpin domain, and both the three-nucleotide and the single-nucleotide loops are unstructured.

As reported above for *c-kit1*, we have monitored representative inter-base distances inside P1 and P2. In P1, we have considered the three W-C base pairs A4–T14, G5–C13 and G6–C12. In P2, the two bases are G2–G19 and G1–G20. For the A–T base pair, the distances are between H3N(T) and N1(A). For the two G–C base-pairs, they are between H1N and N3. In P2, for both G2–G19 and G1–G20, they are between the O6 atoms of the two bases. This choice, rather than the one done for *c-kit1*, was motivated by the different glycosidic angles of the consecutive bases: anti for G2 and G20, and *syn* for G1 and G19. This results into

the N7 atoms of two consecutive bases located on opposite strands. The O6-O6 distances were considered to be less sensitive to the alternation of the glycosidic angles. By contrast, all GQ core guanines in c-kit are anti.

The distance probability density functions (Figure 8, c in the SI)) attest to a larger stability of P2 than P1. It is thus observed that the two O6-O6 distances in P2 are clearly centered around two peaks in a 4.5-4.8 Å interval. In P1, the A4-T14 distance has two peaks. The first corresponds to a standard W-C H-bond distance, but the second at 4.2 Å corresponds to a dissociated pair. For G6-C12, similarly, there are two peaks, the first at 2.0 Å corresponding to a W-C base-pair, and the second at 4.0 Å corresponding to a dissociated base-pair. For G5–C13, a shallow evolution is found in the 2.0 – 4.5 Å range.

Since the cryptic pocket P2 is demonstrated to have a better conformational stability throughout the MD simulation, as clearly shown by the probability density function of distances, drug design efforts should be concentrated on targeting this sub-structure rather than targetting the cryptic pocket P1.

Perspectives on GQs' Druggability: from *c-kit1* to the LTR-III of HIV-1

As mentionned before, *ckit1* is a receptor tyrosine kinase involved in cell proliferation, survival and differentiation. Mutation or overexpression in *ckit1* are associated with cancers, notably gastrointestinal stromal tumprs (GISTs). GQ have been indentified in the promoter region of *ckit1* gene and thus can influence gene expression and can be considered as potential targets for anti-cancer. DoGSite Scorer unraveled two 'cryptic' pockets, P1 and P2, which could serve as 'druggable' sites for potential pharmacological ligands. Figure 8 a) gives a representation of the accessible surface of P1, clearly revealing the existence of a 'groove' for their binding.

Given that GQ have been found also within the HIV LTR-III region, a novel approach to

anti-HIV therapy is provided. We thus decided to consider a relatively similar approach to studythe cryptic pockets that have been found in this regions. We therefore performed a 400 ns CMD on the GQ found on the LTR-III of HIV-1. Here also, DoGSite Scorer identified two cryptic pockets, P1 and P2, of which only P2, that exhibited more limited RMSD variations, could possibly be used as a druggable ligand binding site. Figure 9 gives a representation of the accessible pockets.

On the basis of these results, the design of novel ligands targeting the grooves of cryptic pockets P1 and P2 of *c-kit1*, and of P2 of the LTR-III of HIV-1 becomes possible as well as extending such studies to other G-forming quadruplexes from other oncogenic and retroviral sequences.

The rationale for targeting GQ relies on their unique structural features which make them attractive targets for drug design, as molecules that specifically bind to and stabilize GQ could potentially disrupt biological processes without actually affecting other cellular functions, or causing significant side effects.

Conclusions

We have performed long AMOEBA polarizable MD simulations on two guanine quadruplexes. The first is formed within an GC-rich oncogenic sequence, associated with the *c-kit1* oncogene, and the second is found within a Long Terminal Repeat of the HIV-1 retrovirus. We first focused on a 22-mer oligonucleotide belonging to the *c-kit1* oncogene, the overexpression of which can lead to cancer. This oligonucleotide can fold into guanine quadruplexes, a conformational change occurring during DNA transcription into CD117 mRNA. Understanding the dynamics of this 22-mer oligomer in its quadruplex is essential and was analyzed in detail. Using the AMOEBA polarizable force field coupled to GPU-accelerated unsupervised adaptive sampling, we achieved an aggregated c-kit1 simulation time of 7.5 μs. Most simulation features were found to be fully consistent with those previously ob-

tained with the non-polarizable AMBER[12] or polarizable Drude[24] force fields. However, if the long-duration stability of the GQ core, made of three stacked guanine quadruplexes, and that of the Long-Propeller Loop observed in previous studies were confirmed, this contrasts with the high dynamics of the single-nucleotide loops A5, C9 and of the dinucleotide loop C11-T12, whose mobilities impact the whole dynamics of the *c-kit1* 22-mer. Such mobilities were evaluated in the individual clusters in terms of: a) H-bonding distances in base-pairs A1-T12 and A1-C11, A16-G20, and G17-G20; b) the Two torsion angles, i.e. the glycosidic χ and the γ angles of bases A1, A5, C9, C11 and T12, and c) The π-stacking index between A1 and both bases, C11 and T12 of the dinucleotide loops, and a G base of the nearest guanine tetrad, monitored the evolution in the course of adaptive MD for base-pairs A1-G2, A1-G13, C11-G6, C11-G10, T1-G6, and T12-G10. The tICA analyses revealed 7 distinct clusters transiently stabilized during the MD simulations. Available X-ray structures were identified in the largest detected cluster confirming the GQ core stability. The fact that the identified clusters remain closely related in terms of energy is due to the different conformations of the single and dinucleotides loops. However, all of these clusters share the same stable cryptic pockets that can therefore be considered as potentially druggable sites for drug discovery. GQ's channel simulations using the AMOEBA polarizable force field demonstrated the importance of polarizable water molecules and their potential to effectively replace the effect of cationic ions to stabilize the GQ, preventing it from unfolding as oberved in non-polarizable force field simulations. Our modeling strategy was extended to another model and thus was tested on the LTR-III region of HIV-1. The stability of the G-Quads was also assessed. Well-defined and stable cryptic pockets were also detected. One of them could be exploited as a druggable site for future drug discovery studies.

Technical Appendix

Materials and methods

System construction

The starting structure for *c-kit1* GQ was taken from the crystallography structure in PDB entry *4WO2*.[23] The structure was solvated in a water cubic box with dimensions of 60 Å. To mimic physiological conditions, 150 mM of [KCl] (K^+, Cl^-) was added to the system. The system was then minimized with the AMOEBA polarizable force field[28,50,51] using L-BFGS algorithm until a Root Mean Square (RMS) gradient of 1 kcal/mol was reached.

Conventional Molecular Dynamics

The MD simulations were performed using the recently developed GPU module[39] of the Tinker-HP software,[40] which is part of the Tinker platform.[52] This module efficiently leverages mixed precision,[39] offering a substantial acceleration of simulations. The assembly was equilibrated for 1ns at each of the following temperatures successively: 50, 100, 150, 200, 250, and 300K. The Smooth Particle Mesh Ewald[53] was used with the Ewald and van der Waals cutoffs set to 7 Å and 9 Å, respectively. The MD simulations used the BAOAB-RESPA integrator with a 2 fs outer timestep, a Bussi thermostat, hydrogen-mass repartitioning (HMR), and random initial velocities. Smooth Particle Mesh Ewald (SPME) method was used for handling periodic boundary conditions with a cubic boox of dimensions 60 Å x 60 Å x 60 Å. The production phase was conducted at 300 K for 800 ns for c-kit and 400 ns for HIV-1 LTR-III.

Unsupervised Adaptive Sampling strategy for conformational exploration

Adaptive sampling has proven to be a powerful exploration tool for studying protein folding and dynamics, as well as for exploring a diversity of rare molecular events.[54–57] Recently, we developed a new fully unsupervised Adaptive Sampling strategy that enables the simultaneous use of hundreds of GPU cards.[32] The method divides the simulation procedure into iterations, each composed of a swarm of independent MD simulations run in parallel.

At the beginning of each iteration, some initial structures are selected from the structures sampled in past iterations. The selection procedure is mathematically designed to fully explore a low-dimensional space of slow variables generated by dimensionality reduction algorithms such as PCA and tICA.

From these selected structures, independent MD simulations are run in parallel, potentially combined with other enhanced sampling methods,[58] to generate new structures that can serve as starting points for future iterations.

In this particular case, we first ran an initial 800 ns of cMD where we selected 50 starting point structures using the Adaptive Sampling algorithm in the four-dimensional PCA space. The choice of 50 starting points is heuristic, as we have observed it to be sufficient for efficiently sampling the conformational space of a protein of similar size: the SARS–CoV–2 protein. The number of dimensions for the PCA space is determined based on the fact that the other PCA dimensions have negligible variance contribution below 4 %. Each cMD simulation sampled 15 ns, and these simulations were performed in the NVT ensemble at 300 K. Thus, with a total of 10 iterations, we generated more than 7.5 μs of simulation."

Time-structure Independent Component Analysis

Time-structure Independent Component Analysis (tICA) was introduced into MD as a method to extract slow collective variables.[41,42] tICA solves a variational approach by which

it finds an optimal approximation to the eigenvalues and eigenfunctions of the Markov operator underlying the molecular dynamics data.[41,43,44] As a result, tICA is ideally suited for clustering and to prepare data for Markov models.

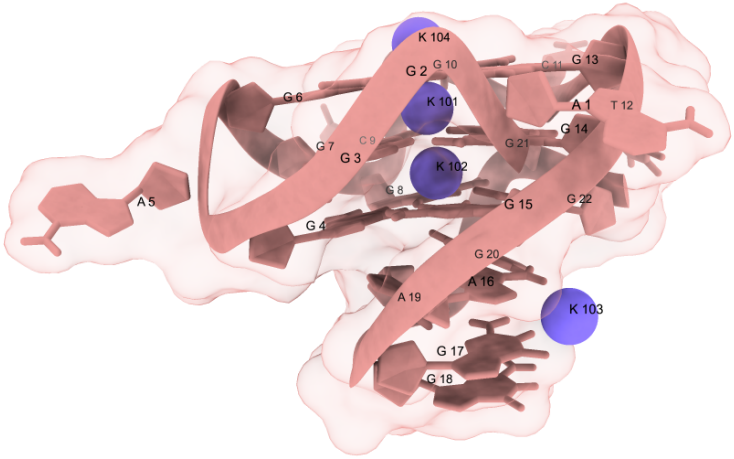
In fact, tICA followed by clustering techniques makes it possible to classify the structural and physico-chemical properties of the structures within an MD trajectory. Additionally, this approach reveals valuable insights into the system's behavior over time.

While tICA shares similarities with PCA, it offers notable advantages for capturing the dynamics of proteins. Unlike PCA, which relies on static and linear properties, tICA aims to create independent components as meaningful as possible to better describe the dynamics of a system.[59] This makes tICA a more suitable choice for our analysis.

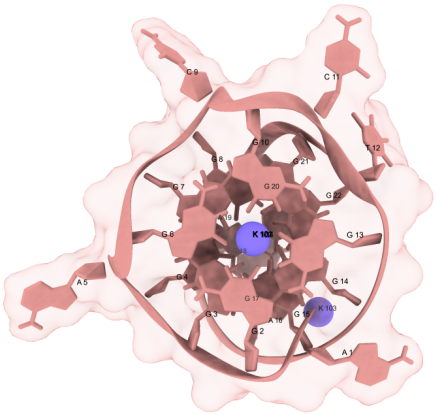
Throughout the simulations, with tICA and clustering techniques we identified distinct clusters for each of the variants analyzed. We organized the final MD trajectories into clusters based on the eigenvalues and eigenfunctions of the covariance matrix.

By employing this classification approach, we can observe similarities between the macro-states of each variant, enabling us to discern common structural or dynamic changes shared among the variants.

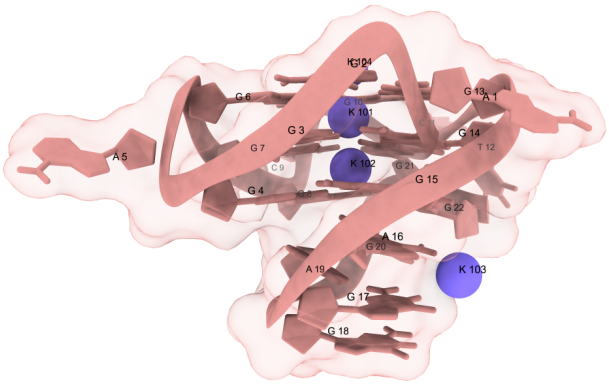
A classification makes it possible to observe the similarities between the macro-states of each of the variants to discern a structural or dynamic change to them.



a)

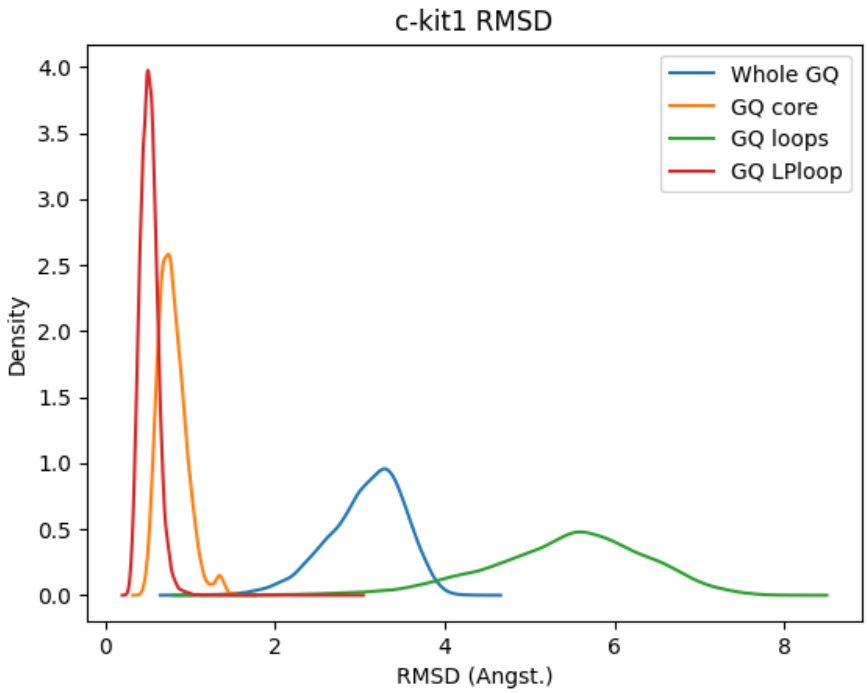


b)

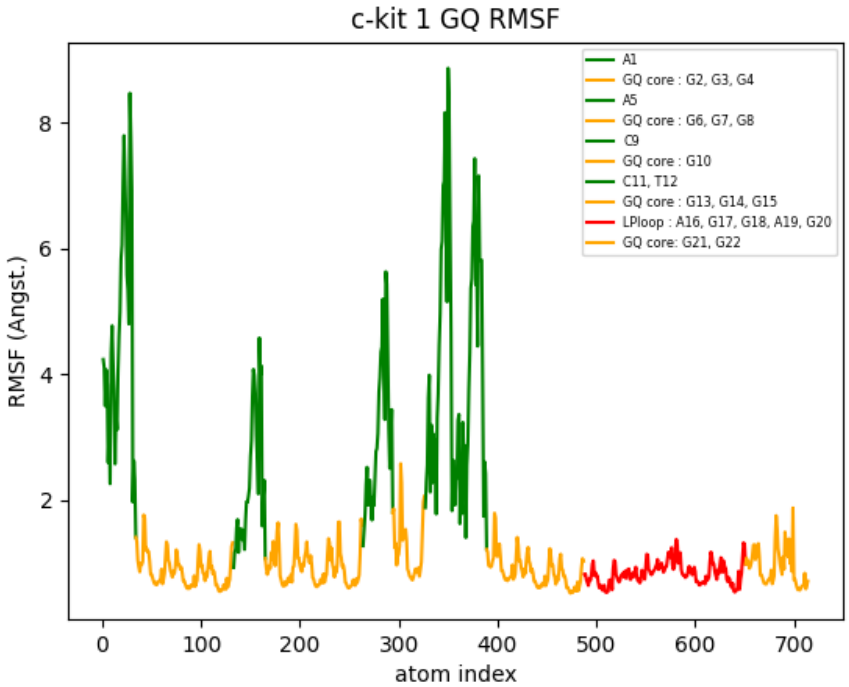


c)

Figure 1: *c-kit1* GQ representation. *c-kit1* is formed by 3 G4 tetrad, one LP loop (A16, G17, G18, A19 and G20), 2 single nucleotide loops (A5 and C9) , one 2 nucleotides loop (C11 and T12) and a free nucleotide (A1). a) General representation showing all the nucleotides of *c-kit1* GQ. b) A top view showing organization of G nucleotides around the ionic channel. c) View from the side for a better visibility of the LP loop.



a)



b)

Figure 2: RMSD and RMSF of *c-kit1*. For the RMSF, the atom indices of nucleotides A1, A5, C9, C11 and T12 are respectively : 0 to 33, 133 to 164, 133 to 164, 133 to 164 and 133 to 164.

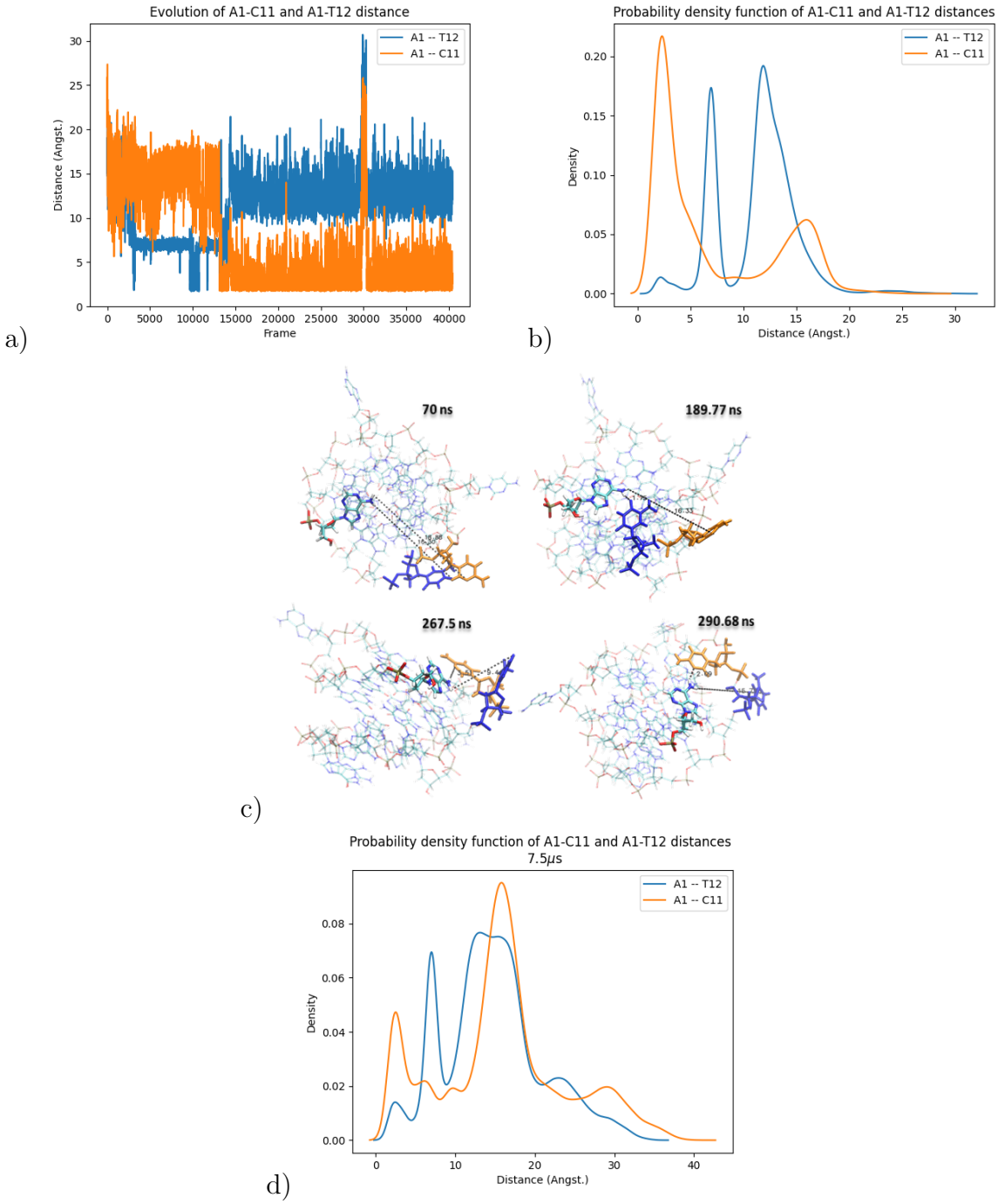


Figure 3: Evolution of the distance between A1(H) and T12(O) shown in blue and between A1(H) and C11(O) shown in orange. 5000 frames correspond to 100 ns simulation time. c) Representation of four structures at four different times of the CMD trajectory showing the evolution of distances A1–T12 and A1–C11.

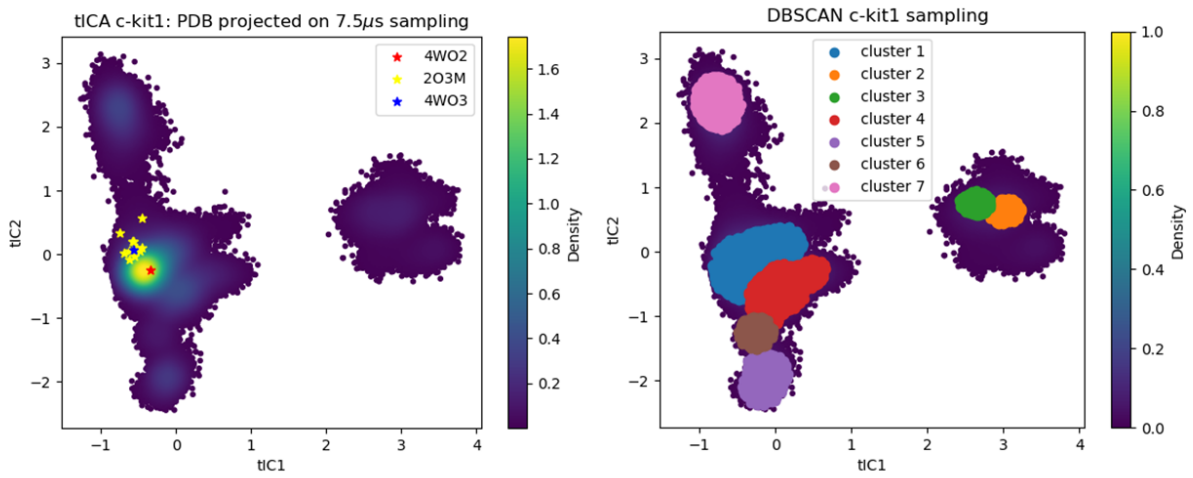


Figure 4: tICA DBSCAN sampling and clustering

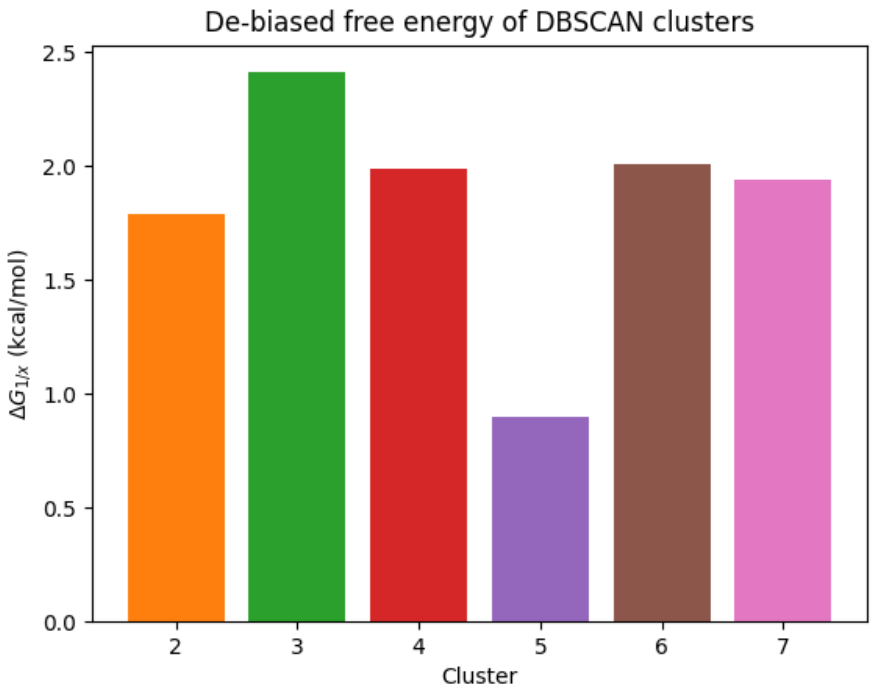


Figure 5: Relative free energies of the DBSCAN 7.5μs tICA clusters, with respect to cluster 1.

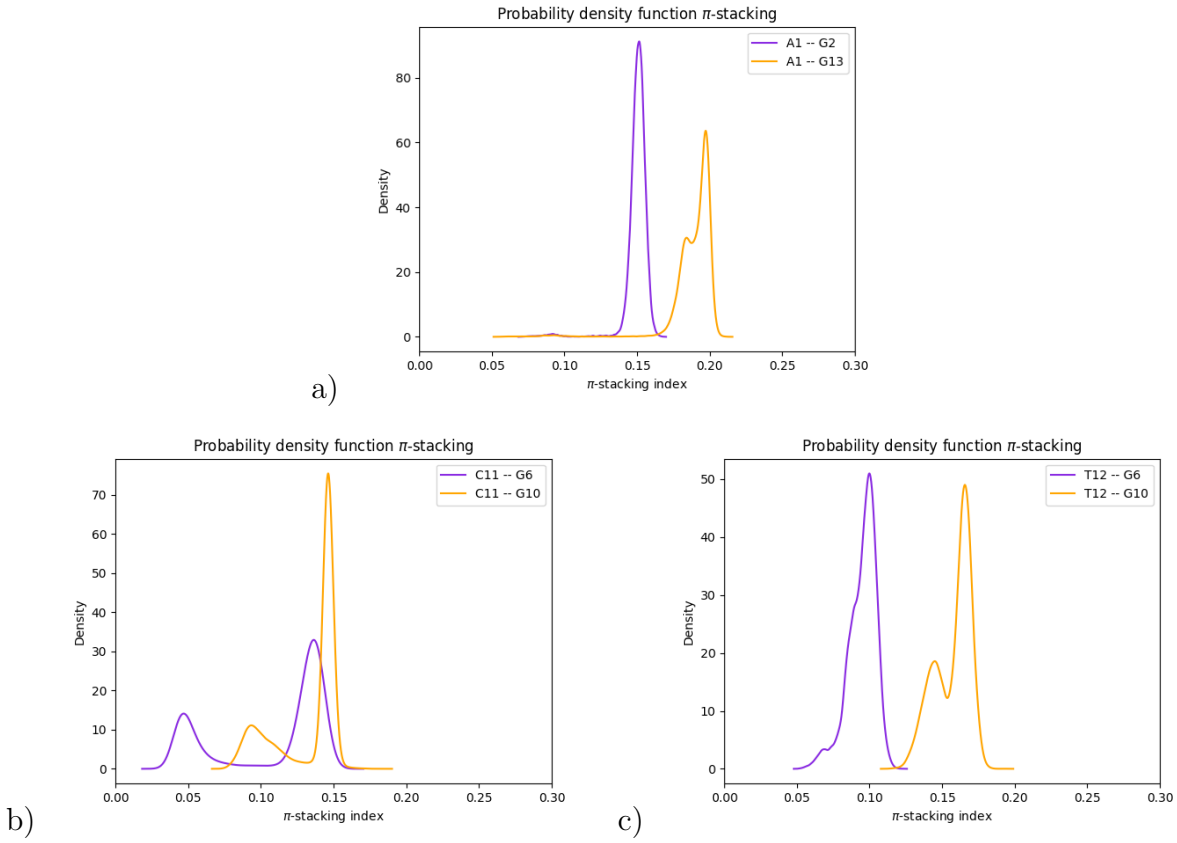


Figure 6: π-π stacking indexes for base pairs A1–G2, A1–G13, C11–G6, C11–G10, T12–G6 and T12–G10 calculated from the adaptive MD trajectories.

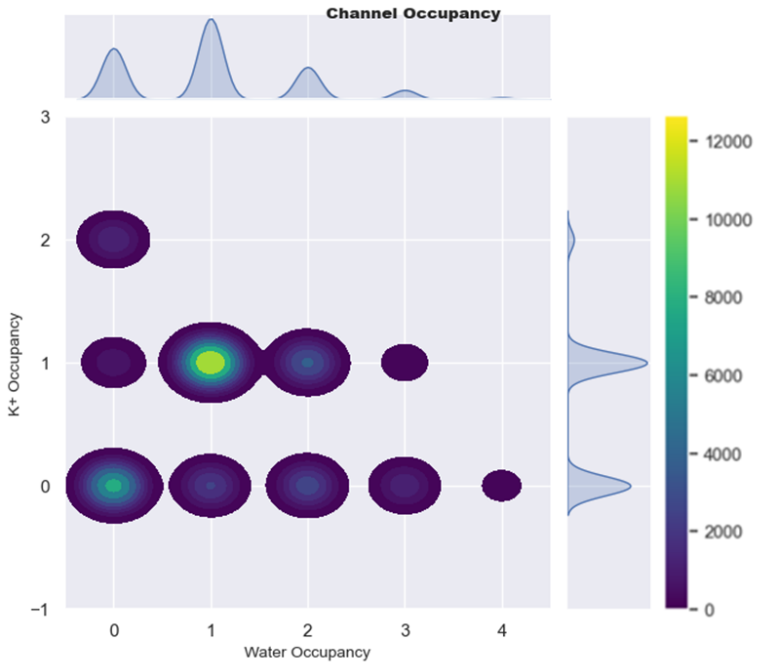
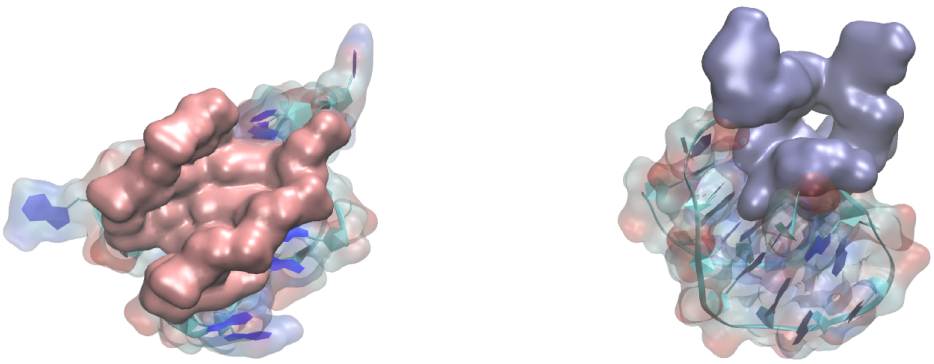
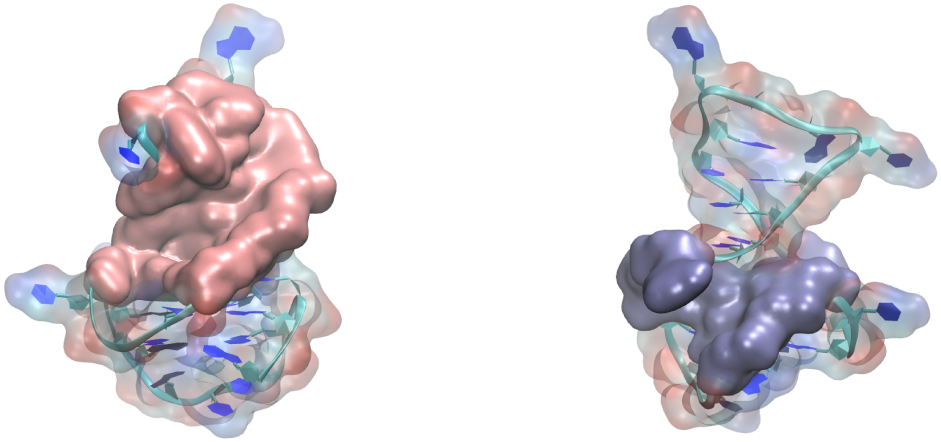


Figure 7: 2D probability density function representation of *c-kit1* GQ channel occupancy throughout the 20 independant trajectories of 20ns each starting from an empty channel. The color bar represents the frame density, from blue to yellow we represent respectively the lowest and the highest density region.



a) b)

Figure 8: Representation of the two cryptic pockets inside *c-kit1* GQ; a) P1 (pink color) formed by the bases A1, G2, G3, G13, G14, G15, A16, and b) P2 (ice blue) formed by bases G7, G8, C9, G10, C11, G21.



a) b)

Figure 9: Representation of the two cryptic pockets inside LTR-III HIV GQ structure c) P1 (pink color) is formed by the hairpin residues and d) P2 (iceblue) is formed by the GQ core bases.

Acknowledgement

DEA thanks funding from the Lebanese National Council for Scientific Research, CNRS-L. FC thanks funding from the French state funds managed by the CalSimLab LABEX and the ANR within the Investissements d'Avenir program (reference ANR11-IDEX-0004-02) and support from the Direction Génerale de l'Armement (DGA) Maîtrise NRBC of the French Ministry of Defense. This work was also made possible thanks to funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 810367), project EMC2, JPP). Computations have been performed at GENCI (IDRIS, Orsay, France and TGCC, Bruyères le Chatel) on grant no A0130712052 (JPP).

Supporting Information Available

Supplementary information is available. It includes: the free energy values of each of the identified tICA clusters; the nucleotides probability density functions; the RMSF of ckit1 GQ components in each of identified tICA clusters.; the c-kit1 GQ RMSD for each of the 20 independent MD simulations starting from an empty ionic channel.

References

(1) Rhodes, D.; Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.* **2015**, *43*, 8627–8637.

(2) 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.

(3) Hardin, C. C.; Watson, T.; Corregan, M.; Bailey, C. Cation-dependent transition be-

tween the quadruplex and Watson-Crick hairpin forms of d(CGCG3GCG). *Biochemistry* **1992**, *31*, 833–841.

(4) Li, W.; Wu, P.; Ohmichi, T.; Sugimoto, N. Characterization and thermodynamic properties of quadruplex/duplex competition. *FEBS letters* **2002**, *526*, 77–81.

(5) Simonsson, T. G-quadruplex DNA structures–variations on a theme. *Biol. Chem.* **2001**, *382*, 621–628.

(6) Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res* **2006**, *34*, 5402–5415.

(7) Wang, Y.; Patel, D. J. Solution structure of the human telomeric repeat d [AG3 (T2AG3) 3] G-tetraplex. *Structure* **1993**, *1*, 263–282.

(8) Horvath, M. P.; Schultz, S. C. DNA G-quartets in a 1.86 Å resolution structure of an Oxytricha nova telomeric protein-DNA complex. *Journal of molecular biology* **2001**, *310*, 367–377.

(9) Parkinson, G. N.; Lee, M. P.; Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **2002**, *417*, 876–880.

(10) Georgiades, S.; AbdKarim, N.; Suntharalingam, K.; Vilar, R. Interaction of Metal Complexes with G-Quadruplex DNA. *Angewandte Chemie International Edition* **2010**, *49*, 4020–4034.

(11) Phan, A. T.; Kuryavyi, V.; Burge, S.; Neidle, S.; Patel, D. J. Structure of an Unprecedented G-Quadruplex Scaffold in the Human c-kit Promoter. *J. Am. Chem. Soc.* **2007**, *129*, 4386–4392.

(12) Islam, B.; Stadlbauer, P.; Krepl, M.; Koca, J.; Neidle, S.; Haider, S.; Sponer, J. Extended molecular dynamics of a c-kit promoter quadruplex. *Nucleic Acids Res* **2015**, *43*, 8673–8693.

(13) Liu, L.; Bailey, S. M.; Okuka, M.; Muñoz, P.; Li, C.; Zhou, L.; Wu, C.; Czerwiec, E.; Sandler, L.; Seyfang, A.; Blasco, M. A.; Keefe, D. L. Telomere lengthening early in development. *Nat Cell Biol.* **2007**, *9*, 1436–41.

(14) Brooks, T. A.; Hurley, L. H. The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics. *Nature Reviews Cancer* **2009**, *9*, 849–861.

(15) Heinrich, M. C. et al. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *J Clin Oncol.* **2003**, *21*, 4341–4349.

(16) Huppert, J. L.; Bugaut, A.; Kumari, S.; Balasubramanian, S. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res* **2008**, *36*, 6260–6268.

(17) Frolov, A.; Chahwan, S.; Ochs, M.; Arnoletti, J. P.; Pan, Z.-Z.; Favorova, O.; Fletcher, J.; von Mehren, M.; Eisenberg, B.; Godwin, A. K. Response Markers and the Molecular Mechanisms of Action of Gleevec in Gastrointestinal Stromal Tumors. *Mol Cancer Ther.* **2003**, *2*, 699–709.

(18) Foster, B. M.; Zaidi, D.; Young, T. R.; Mobley, M. E.; Kerr, B. A. CD117/c-kit in Cancer Stem Cell-Mediated Progression and Therapeutic Resistance. *biomedicines 6*.

(19) Babaei, M. A.; Kamalidehghan, B.; Saleem, M.; Huri, H. Z.; Ahmadipour, F. Receptor tyrosine kinase (c-Kit) inhibitors: a potential therapeutic target in cancer cells. *Drug Des Devel Ther* **2016**, *10*, 2443–2459.

(20) Bhamidipati, P. K.; Kantarjian, H.; Cortes, J.; Cornelison, A. M.; Jabbour, E. Management of imatinib-resistant patients with chronic myeloid leukemia. *Therapeutic Advances in Hematology* **2013**, *4*, 103–117.

(21) Rankin, S.; Reszka, A. P.; Huppert, J.; Zloh, M.; Parkinson, G. N.; Todd, A. K.; Ladame, S.; Balasubramanian, S.; Neidle, S. Putative DNA Quadruplex Formation within the Human c-kit Oncogene. *J. Am. Chem. Soc.* **2005**, *127*, 10584–10589.

(22) Wei, D.; Husby, J.; Neidle, S. Flexibility and structural conservation in a c-KIT G-quadruplex. *Nucleic Acids Res* **2015**, *43*, 629–644.

(23) Wei, D.; Parkinson, G. N.; Reszka, A. P.; Neidle, S. Crystal structure of a c-kit promoter quadruplex reveals the structural role of metal ions and water molecules in maintaining loop conformation. *Nucleic Acids Res* **2012**, *40*, 4691–4700.

(24) Salsbury, A. M.; Lemkul, J. A. Molecular Dynamics Simulations of the c-kit 1 Promoter G-Quadruplex: Importance of Electronic Polarization on Stability and Cooperative Ion Binding. *J Phys Chem B.* **2019**, *123*, 148–159.

(25) Perrone, R.; Nadai, M.; Frasson, I.; Poe, J. A.; Butovskaya, E.; Smithgall, T. E.; Palumbo, M.; Palù, G.; Richter, S. N. A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter. *J Med Chem.* **2013**, *19*, 6521–6530.

(26) Piekna-Przybylska, D.; Sullivan, M. A.; Sharma, G.; Bambara, R. A. U3 region in the HIV-1 genome adopts a G-quadruplex structure in its RNA and DNA sequence. *Biochemistry.* **2014**, *16*, 2581–2593.

(27) Amrane, S.; Kerkour, A.; Bedrat, A.; Vialet, B.; Andreola, M.-L.; Mergny, J.-L. Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development. *J Am Chem Soc.* **2014**, *14*, 5249–5252.

(28) Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **2018**, *14*, 2084–2108.

(29) Shi, Y.; Ren, P.; Schnieders, M.; Piquemal, J.-P. *Reviews in Computational Chemistry Volume 28*; John Wiley and Sons, Ltd, 2015; Chapter 2, pp 51–86.

(30) Melcr, J.; Piquemal, J.-P. Accurate biomolecular simulations account for electronic polarization. *Front. Mol. Biosci.* **2019**, *6*, 143.

(31) Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Ann. Rev. Biophys.* **2019**, *48*, 371–394.

(32) Jaffrelot-Inizan, T.; Célerse, F.; Adjoua, O.; Ahdab, D. E.; Jolly, L.-H.; Montes, M.; Lagarde, N.; Ren, P.; Lagardère, L.; Monmarché, P.; Piquemal, J.-P. High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. *Chem. Sci.* **2021**, *12*, 4889–4907.

(33) El Ahdab, D.; Lagardère, L.; Jaffrelot Inizan, T.; Célerse, F.; Liua, C.; Adjoua, O.; Jolly, L.-H.; Gresh, N.; Hobaika, Z.; Ren, P.; Maroun, R. G.; Piquemal, J.-P. Interfacial Water Many-Body Effects Drive Structural Dynamics and Allosteric Interactions in SARS-CoV-2 Main Protease Dimerization Interface. *J. Phys. Chem. Lett.* **2021**, *12*, 6218–6226.

(34) El Khoury, L. et al. Computationally driven discovery of SARS-CoV-2 Mpro inhibitors: from design to experimental validation. *Chem. Sci.* **2022**, *13*, 3674–3687.

(35) Blazhynska, M.; Lagardère, L.; Liu, C.; Adjoua, O.; Ren, P.; Piquemal, J.-P. Water–glycan interactions drive the SARS-CoV-2 spike dynamics: insights into glycan-gate control and camouflage mechanisms. *Chem. Sci.* **2024**, DOI: 10.1039/D4SC04364B.

(36) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. Anisotropic, Polarizable Molecular Mechanics Studies of Inter- and Intramolecular Interactions and Ligand-Macromolecule Complexes. A Bottom-Up Strategy. *Journal of Chemical Theory and Computation* **2007**, *3*, 1960–1986, PMID: 18978934.

(37) Gresh, N.; Naseem-Khan, S.; Lagardère, L.; Piquemal, J.-P.; Sponer, J. E.; Sponer, J. Channeling through Two Stacked Guanine Quartets of One and Two Alkali Cations in the Li+, Na+, K+, and Rb+ Series. Assessment of the Accuracy of the SIBFA

Anisotropic Polarizable Molecular Mechanics Potential. *J. Phys. Chem. B* **2017**, *121*, 3997—-4014.

(38) 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. Theory Comput.* **2015**, *11*, 4972–4991.

(39) Adjoua, O.; Lagardère, L.; Jolly, L.-H.; Durocher, A.; Very, T.; Dupays, I.; Wang, Z.; Inizan, T. J.; Célerse, F.; Ren, P.; Ponder, J. W.; Piquemal, J.-P. Tinker-HP: Accelerating Molecular Dynamics Simulations of Large Complex Systems with Advanced Point Dipole Polarizable Force Fields Using GPUs and Multi-GPU Systems. *Journal of Chemical Theory and Computation* **2021**, *17*, 2034–2053, PMID: 33755446.

(40) Lagardère, L.; Jolly, L.-H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J.; Gresh, N.; Maday, Y.; Ren, P. Y.; Ponder, J. W.; Piquemal, J.-P. Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.* **2018**, *9*, 956–972.

(41) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*.

(42) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(43) Noé, F.; Nüske, F. A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems. *Multiscale Model. Simul.* **2013**, *11*, 635–655.

(44) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J Chem Theory Comput.* **2014**, *8*, 1739–52.

(45) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; others A density-based algorithm for discovering clusters in large spatial databases with noise. kdd. 1996; pp 226–231.

(46) Branduardi, D.; Gervasio, F. L.; Cavalli, A.; Recanatini, M.; Parrinello, M. The role of the peripheral anionic site and cation-π interactions in the ligand penetration of the human AChE gorge. *J. Am. Chem. Soc.* **2005**, *127*, 9147–9155.

(47) Lipfert, J.; Doniach, S.; Das, R.; Herschlag, D. Understanding nucleic acid–ion interactions. *Annual review of biochemistry* **2014**, *83*, 813–841.

(48) Ida, R.; Wu, G. Direct NMR detection of alkali metal ions bound to G-quadruplex DNA. *Journal of the American Chemical Society* **2008**, *130*, 3590–3602.

(49) Volkamer, A.; Kuhn, D.; Rippmann, F.; Rarey, M. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics.* **2012**, *28*, 2074–2075.

(50) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(51) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4046–4063.

(52) Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory and Comput.* **2018**, *14*, 5273–5289.

(53) Lagardère, L.; Lipparini, F.; Polack, ; Stamm, B.; Cancès, ; Schnieders, M.; Ren, P.; Maday, Y.; Piquemal, J.-P. Scalable Evaluation of Polarization Energy and Associated

Forces in Polarizable Molecular Dynamics: II. Toward Massively Parallel Computations Using Smooth Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2015**, *11*, 2589–2599, PMID: 26575557.

(54) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory. Comput.* **2010**, *6*, 787–794.

(55) Zimmerman, M. I.; Porter, J. R.; Sun, X.; Silva, R. R.; Bowman, G. R. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *J. Chem. Theory. Comput.* **2018**, *14*, 5459–5475.

(56) Betz, R. M.; Dror, R. O. How Effectively Can Adaptive Sampling Methods Capture Spontaneous Ligand Binding? *J. Chem. Theory. Comput.* **2019**, *15*, 2053–2063.

(57) Hruska, E.; Abella, H. R.; Nüske, F.; Kavraki, L. E.; Clementi, C. Quantitative comparison of adaptive sampling methods for protein dynamics. *J. Chem. Phys.* **2018**, *149*, 244119.

(58) Célerse, F.; Inizan, T. J.; Lagardère, L.; Adjoua, O.; Monmarché, P.; Miao, Y.; Derat, E.; Piquemal, J.-P. An Efficient Gaussian-Accelerated Molecular Dynamics (GaMD) Multilevel Enhanced Sampling Strategy: Application to Polarizable Force Fields Simulations of Large Biological Systems. *Journal of Chemical Theory and Computation* **2022**, *18*, 968–977, PMID: 35080892.

(59) Naritomi, Y.; Fuchigami, S. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.* **2013**, *139*, 215102.