To probe the binding pathway of a selective compound (D089-0563) to c-MYC Pu24 G-quadruplex using free ligand binding simulations and Markov state model analysis

Brian Chen[1,#], Griffin Fountain[1,#], Holli-Joi Sullivan[1], Nicholas Paradis[1] and Chun Wu[1]*

[#]These authors contributed equally

[1] Rowan University, College of Science and Mathematics, Glassboro, NJ, 08028 USA

* To whom correspondence should be addressed. Email: wuc@rowan.edu

Abstract:

D089-0563 is a highly promising anti-cancer compound that selectively binds the transcription-silencing G-quadruplex element (Pu27) at the promoter region of the human c-MYC oncogene; however, its binding mechanism remains elusive. The structure of Pu27 is not available due to its polymorphism, but the G-quadruplex structures of its two shorter derivatives in complex with a ligand (Pu24/Phen-DC3 and Pu22/DC-34) are available and show significant structural variance as well as different ligand binding patterns in the 3' region. Because D089-0563 shares the same scaffold as DC34 while having a significantly different scaffold from Phen-DC3, we picked Pu24 instead of Pu22 for this study in order to gain additional ligand binding insight. Using free ligand molecular dynamics binding simulations (33 μs), we probed the binding of D089-0563 to Pu24. Our clustering analysis identified three binding modes (top, side, and bottom) and subsequent MMPBSA binding energy analysis identified the top mode as the most thermodynamically stable. Our Markov State Model (MSM) analysis revealed that there are three parallel pathways for D089-0563 to the top mode from unbound state and that the ligand binding follows the conformational selection mechanism. Combining our predicted complex structures with the two experimental structures, it is evident that structural differences in the 3' region between Pu24 and Pu22 lead to different binding behaviors despite having similar ligands; this also explains the different promoter activity caused by the two G-quadruplex sequences observed in a recent synthetic biology study. Based on interaction insights, 625 D089-0563 derivatives were designed and docked; 59 of these showed slightly improved docking scores.

Introduction:

A G-quadruplex consists of four guanine base pairs (G-tetrads) stabilized by Hoogsteen hydrogen binding and pi-pi stacking interactions[1].The G-tetrads form a planar arrangement within the G-quadruplex structure and are further stabilized by the presence of monovalent cations, such as K^+ and Na^+, which interact with guanine O6 carbonyl groups[2]. G-quadruplexes have been found in the promoter regions of numerous genes, including various oncogenes such as c-MYC, c-KIT, VEGF, and KRAS [3-5]. Investigations of new quadruplex-binding ligands that can stabilize the quadruplex structure have gained interest over the past two decades [3-5,6, 7]. Specifically, ligand-binding induced stabilization of the promoter G-quadruplex is being explored as a cancer therapy because of its ability to downregulate oncogenes expression [8-10].

Overexpression of the c-MYC oncogene has gained a lot of attention due to its common genetic aberrations found in various types of human cancer cells, including breast [11], prostate[12], lung [13], cervical [14], colon [15], small-cell lung cancer [16], and lymphoma [17, 18]. The c-MYC gene encodes transcription factors, which regulate gene expression in many important biological processes such as cell growth, apoptosis and proliferation [19]. Within the MYC promoter region, an important element, termed the nuclease-hypersensitivity element III_1 (NHE III_1), is required for 80-95% of c-MYC transcription [20-22]. Pu27, the major purine rich strand of NHE III_1, [8, 23, 24] forms G-quadruplex structures and ligand-binding induced stabilization of the c-MYC G-quadruplex has been shown to downregulate c-MYC expression as a cancer therapy [8, 9]. In one study, suppression of MYC expression was observed when a Burkitt's lymphoma cell line was treated with TMPyP4, which stabilized the G-quadruplex[9]. Two additional studies have also shown that Quindoline derivatives stabilize the c-MYC G-quadruplex and reduce expression of c-MYC in cancer cells [25, 26]. Therefore, the use of small molecules to stabilize the c-MYC G-quadruplex and consequently decreasing MYC expression is an attractive anti-cancer therapeutic approach.

In an attempt to identify a new class of MYC G-quadruplex stabilizing ligands, Felsenstein et al [27] employed a small molecule microarray to screen 20,000 compounds from ChemBridge and ChemDiv repositories, yielding compound 1 (ChemDiv No. D089-0563): a crescent-shaped structure containing a

G-quadruplex binding Disubstituted Benzofuran scaffold. Measurement of the binding affinity of D089-0563 (hereafter referred as DBD1) by surface plasmon resonance gave a K_d value of 4.5 ± 1.4 μM, which is sufficient to elicit a biological response[10]. DBD1, a G-quadruplex stabilizing ligand, also exhibited the rare ability to selectively inhibit c-MYC gene expression through stabilization of Pu27 in the c-MYC promoter region and induced apoptosis in cancer cell lines while having minimal toxicity on normal peripheral blood mononucleocytes[10, 28]. It is rare for a G-quadruplex ligand to specifically target a single G-quadruplex, but DBD1 has been shown to selectively target the c-MYC G-quadruplex[29]. Although the high resolution complex structure of DBD1 with the c-MYC G-quadruplex of Pu27 has not been obtained, the NMR structures of its two short derivatives, Pu24 (PDB: 2MGN) and MYC22-G14T/G23T (Pu22, PDB: 5W77) in complex with different ligands (Pu24/Phen-DC3 and Pu22/DC-34) have been solved which offer some insight (**Figure 1 and Figure S1**). The structures of both G-quadruplexes are very similar except that while the extended 3'-end of Pu24 is connected to the bottom G-triad plane and restricts bottom-binding of Phen-DC3 to Pu24, the truncated 3'-end of Pu22 allows bottom-binding of DC-34 to Pu22 (**Figure 1 and S1**).

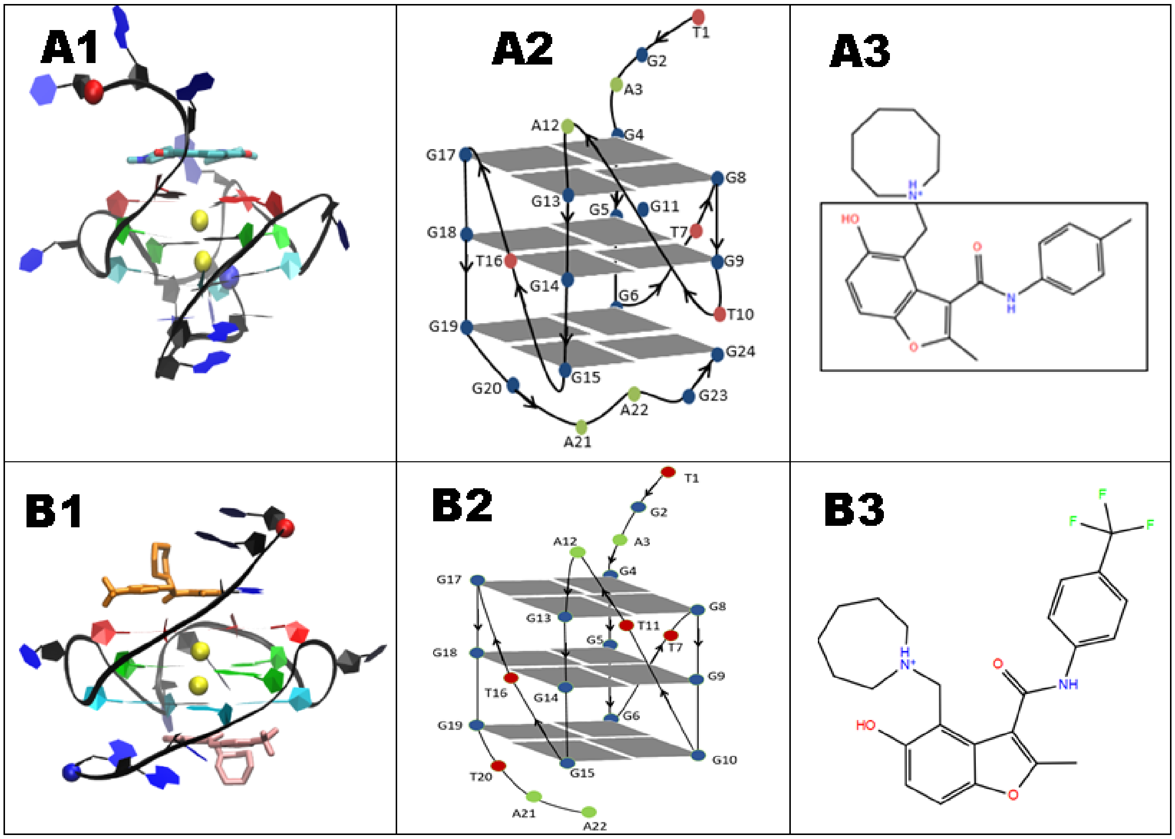


Figure 1. Structure of Pu24 and Pu22 G-quadruplexes and ligands. Side view **(A1)** and cartoon representation **(A2)** of the Pu24 G-quadruplex (PDB ID: 2MGN) and the 2D structure of DBD1(**A3**). Side view **(B1)** and cartoon representation **(B2)** of the Pu22 G-quadruplex (PDB ID: 5W77) and the 2D structure of DC-34 (**B3**). The G-tetrads are highlighted in red, green and cyan. The 5' and 3' ends are represented by a red and blue ball, respectively. K+ cations are represented by yellow balls. The planar core region of the DBD1 structure is shown by a black box.

Additionally, when comparing Pu22 and Pu24 to Pu27, the original guanine tracts from Pu27 are more conserved in the structure of Pu24 as compared to that of Pu22. A study by Sengupta et al. showed that Pu24 is also biologically more similar to Pu27 and that the Pu24 conformation helps initiate transcription[30]. While Phen-DC3, the ligand in complex with Pu24, is much larger and significantly different in structure than DBD1, DC-34, the ligand in complex with the Pu22, is very similar to DBD1, simply replacing a methyl with a trifluoromethyl. Therefore, there is more information to be gained from studying DBD1 in conjunction with Pu24 rather than Pu22. Thus, to gain more insight in this study, we modeled the G-quadruplex of Pu24 for our free ligand binding simulations of DBD1 (**Figure S2**) and compared our simulation results with both experimental structures.

MD simulations have become a powerful and invaluable tool for discovering details on the G-quadruplex over the past two decades [31-33]. Molecular modeling techniques are a widely used method to understand the binding of small molecules to G-quadruplexes and provide a good structure model of the complex. Several studies have used molecular modeling techniques to explore the anti-cancer effects of small molecule stabilization of the c-MYC G-quadruplex [25, 34-38]. However, the binding mechanisms for ligands to the G-quadruplexes from an unbound state are relatively unexplored. Two major theories to describe the dynamic binding mechanisms are the induced-fit mechanism, which proposes that the ligand binding to its target causes a conformational change in the receptor to create a more optimal binding topology [39], and the conformational selection mechanism, which proposes that a ligand binds to a specific conformation of a target and shifts the population of this favorable conformation accordingly [40]. While many studies have been carried out regarding the binding of small molecules to the c-MYC G-quadruplex [25, 29, 35-38, 41], there are currently no studies about the binding mechanism or the binding pathway of DBD1 to the c-MYC G-quadruplex.

Markov State Models (MSM) can be built from MD simulation data and are a comprehensive statistical approach used to create understandable yet high-resolution models of the intrinsic kinetics of a system[42]. MSMs construct a network model with a series of N states and perform calculations based on the transition rates between these states[42]. While traditionally used to study protein folding, MSMs have also been used to study the binding pathways of protein-ligand systems, such as the LAO protein, to determine the primary binding mechanism[43, 44]. However, the application of MSMs in studying the binding pathways of G-quadruplex-ligand systems are relatively unexplored.

In this study, we employ extensive free ligand binding MD simulations (33 μs) to probe the binding of DBD1 to a c-MYC G-quadruplex derivative, Pu24 (PDB: 2MGN). Subsequent k-means clustering analysis of the MD simulation data is evaluated to determine the major binding modes. We compare our MD binding modes with the two relative experimental structures and the implications are discussed. A Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) analysis details the energetics of the major binding modes. MSM analysis examines the binding pathway and kinetic rate information. Order parameters are calculated to characterize the representative trajectories. Geometry and fluctuation of the G-quadruplex structures with and without ligands are examined. Based on interaction insights, D089-0563 derivatives were designed and docked; several derivatives showed improved docking scores.

Methods

C-MYC G-quadruplex structure selection: The high-resolution structure of c-MYC Pu27 is not yet available. It is the dynamic and polymorphic conformation which limits the structural determination of Pu27.[30] Specially, the structural determination is challenging because Pu27's 27-mer sequence consists of six guanine tracts (GI-GVI) and exhibits a large disparity in the length of the intervening loops which allows for shuffling between the G-tracts of quadruplex loops. [45-48] In order to resolve this issue, Pu27 was truncated into a shorter strand and modified by base substitution and/or insertions. The NMR structures of Pu27's two shortened derivatives, Pu24 (PDB: 2MGN) and MYC22-G14T/G23T (Pu22, PDB: 5W77), were solved in 2014 and 2018 [24, 29]. Both derivatives adopt a similar G-quadruplex scaffold as the Pu27 structure (**Table S1 and Figure S1**). Pu24, with only GI and GVI tracts removed and a G10T substitution has been shown to also have a relatively clean NMR spectra.[49] The second structure, Pu22, is derived from a G14T and G23T substitution of MYC22[8]. MYC22 is an extended version of Pu18, a structure that only contained the necessary guanine tracts for G-quadruplex formation[8]. We chose to use the Pu24 structure in our study, because the 24 nucleotide sequence is likely more closely related to the 27 nucleotide wild type sequence (Pu27) than the 22 nucleotide sequence (Pu22 G14T/G27T). Further corroboration of this conclusion was obtained from a study by Sengupta et al, where the c-MYC promoter activity of Pu24 was shown to be significantly greater than that of Pu22 across multiple cancer cell lines while maintaining similar levels of promoter activity to that of the wild-type Pu27 [30]. Nonetheless, the Pu22 structure provides an excellent reference to validate our structure predictions made on our MD simulations of Pu24 for two reasons: (1) the Pu22 structure is very similar to the structure of Pu24 and (2) the ligand in complex with

Pu22, DC-34, is very similar to DBD1 where a methyl is replaced with a trifluoromethyl **(Figure S1)** [29]. To highlight their structural similarities and differences we constructed cartoon representations of the Pu24 and Pu22 G-quadruplexes (**Figure S1**). Both G-quadruplexes are similar in the tetranucleotide loop at the 5'-end (5'-TGAG) and most nucleotide positions are conserved prior to reaching the 22nd nucleotide base (A22). Where the most significant difference between Pu24 and Pu22 is the additional two nucleotides on the 3'-end of Pu24 (GG-3') and their absence on Pu22. The extended 3'-end, connected to the bottom G-tetrad plane of Pu24, restricts bottom-binding of Phen-DC3 to Pu24; the truncated 3'-end of Pu22 allows bottom-binding of DC-34 to Pu22. Other differences between Pu24 and Pu22 include switching of the 10th nucleotide (T and G for Pu24 and Pu22, respectively) and the 20th nucleotide (G and T for Pu24 and Pu22, respectively). Regardless, the nucleotide sequences of the bottom loop of Pu24 and Pu22 highlight their key difference and is of significant focus in our study.

MD simulation systems: A total of 3 systems were constructed: one ligand only system, one DNA only system and an unbound DNA-ligand system. **(Table 1).**

Table 1. Molecular dynamics simulations.

System ID	DNA	No. of ligand	No. of run	Drug Initial State	NPT eq. (ns)	NVT (ns)	Total time (μs)
1	N/A	1 DBD1	4	N/A	1	1000	4
2	Pu24	0	4	N/A	1	1000	4
3	Pu24	1 DBD1	33	Unbound	1	1000	33

Each system was solvated in a water box of truncated octahedron with 10 Å water buffer plus Cl- or K+ as counter ions to neutralize the system and 0.15 M KCl salt concentration. A refined OL15 version of the AMBER nucleic acid force including corrections of several backbone torsion angle parameters (i.e. parm99bsc0 + χ_{OL4} + ε/ζ_{OL1} + β_{OL1})[50-53] was applied to represent the DNA fragment.   TIP3P water model was used to represent water molecules, and the K^+ model developed by Cheatham group was used to represent the K^+ ions.[54, 55] The force field for DBD1 molecule was obtained using standard AMBER protocol: the molecular electrostatic potential (MEP) of the DBD1 molecule was calculated at the HF/6-31G* level after its geometry optimization at the same theory level; then MEP was used to determine the partial charges of DBD1 atoms using Restrained Electrostatic Potential/RESP method with two stage fitting; and other force field parameters were taken from the AMBER GAFF2 force field. [56, 57]

MD Simulation protocols: The simulations for each system was ran using AMBER 16 simulation package.[56] The simulation protocols followed our early studies [58-60] which are briefly described here. The starting points of the MD simulations for DBD1 to the G-quadruplex involved two different initial starting unbound points, top and bottom, with a separation of ~15 Å (**Figure S2**). Each unbound DNA-ligand system underwent an additional 1000 ps pre-run at 500K to ensure that the position and orientation of the free ligand was randomized before a production run at 300K; during this pre-run, the receptors position remained fixed. Thirty-three independent runs at 300K were carried out using random initial velocities. A run at 300 K, included a short 1.0 ns molecular dynamics in the NPT ensemble mode (constant pressure and temperature) to equilibrate the system density and production dynamics in the equivalent NVT ensemble mode (constant volume and temperature). SHAKE was applied to constrain all bonds connecting hydrogen atoms, enabling a 2.0 fs time step in the simulations[61]. The particle-mesh Ewald method was used to treat long-range electrostatic interactions under periodic boundary conditions (charge grid spacing of ~1.0 Å, the fourth order of the B-spline charge interpolation; and direct sum tolerance of 10^{-5})[62]. The cut off distance for short-range non-bonded interactions was 10 Å, with the long-range van der Waals interactions based on a uniform density approximation. To reduce the computation, non-bonded forces were calculated using a two-stage RESPA approach where the short range forces were updated every step and the long range forces were updated every two steps[63]. Temperature was controlled using the Langevin thermostat with a coupling constant of 2.0 ps. The trajectories were saved at 50.0 ps intervals for analysis.

Convergence of simulations: The distribution of DBD1 over the course of all 33 binding simulations is presented in the supporting document where DBD1 is represented by a single atom (**Figure S3**). From the top view and side view, it was evident the DNA G-quadruplex surface was well sampled by the ligand, suggesting that a good position sampling was been achieved by our simulation protocol. Using the first frame of each trajectory as the reference frame, the trajectory was aligned on the heavy atoms of both DNA backbone and the ligand. After alignment, the Root Mean Square Deviation (RMSD) of the heavy atoms of both DNA backbone **(Figure S4)** and the ligand **(Figure S5)** was calculated for all runs of the free ligand binding simulations. Atom contacts between the DNA structure and the drug molecule were calculated using an atom-to-atom distance cutoff of 3.0 Å (**Figure S6**). Flat RMSDs and atom contacts were observed after 250 ns in most of the trajectories, indicating the convergence of the binding simulations was achieved. The last snapshot of all 33 trajectories are shown in **Figure S7**. The convergence of the Markov state model for the side, top and bottom binding states is shown in **Figure S9 and S10**. In **Figure** S9, each of the thirty-three 1001.0 ns MD simulations was divided into three blocks (A, B and C) constituting 333.3 ns of simulation time. Calculation of the implied timescales for each block showed all three bound states converging (i.e. flattening of data curve) around lagtime 250 ns in each block. These results were similarly observed in the implied timescales of the overall trajectory data (**Figure S10**), where all three bound states also converge at around a lagtime of 250 ns.

G-quadruplex parameters: Following our previous studies[64, 65], the rise, helical rise (H-rise), and helical twist (H-twist) were calculated to characterize the geometry of the G-quadruplex core **(Table 2)**. Rise was defined as the distance (Å) between the guanine base centers, excluding the sugar phosphate backbone, of the lower G4 layer to the guanine base centers of the G4 layer above. H-rise was defined as the projection of the G-quartet rise with respect to the Z-axis. We defined the Z-axis as the vertical axis through the two K+ cations within the G4 ion channel. H-twist was defined as the rotational angle of the bases with respect to the Z-axis. The H-twist rotation was calculated from the bottom, where the layer above would be used as a reference, measuring the degree of clockwise rotation required of the lower layer to align with the position of the reference layer.

Potassium Ion Position: Electrostatic interactions between the K+ cations and the partially negative oxygen's of each guanine core residue plays an integral role in stabilizing the G-quadruplex structure[66]. The DNA apo form was used to analyze the position of each K+ cation in reference to the surrounding DNA G-quartets. This analysis measured the distance between the oxygen on each G-quartet residue, the distance from this residue to the nearest K^+ cation, as well as the distance between the oxygen of each residue to the oxygen of the nearest residue (**Figure 10**). The oxygen-oxygen and oxygen-potassium distances for the apo form and the top, bottom, and side major binding poses can be found in the supporting document (**Table S2-S3**).

Featurization and Clustering: 33 trajectories (1000 ns each) of the DNA-ligand system were combined into one trajectory. Using VMD, all frames in which there were less than 13 atom contacts, at a distance less than 3Å, between the G-quadruplex and the ligand were separated as the unbound state [67]. The trajectory was then superimposed based on the nucleic backbone only using MDtraj, then calculations for RMSD of the ligand heavy atoms without least square fit as well as center of mass (COM) of the ligand heavy atoms were performed [68]. The ligand was not included in the structure alignment, because of its large position, orientation and conformation fluctuation. The DNA was not included in the RMSD calculation for clustering, because its structure is relatively rigid, leading small RMSD fluctuation. K-means clustering on four features (i.e. the RMSD and each of the three coordinates of the COM) performed using scikit-learn, was then used to classify the remaining frames into various states [69]. Clustering was performed for K between 2 and 30 inclusively, using the silhouette index as the metric for similarity of clusters [69, 70]. It was determined that K=4 had the greatest silhouette index and the most representative frame for each cluster was determined by calculating the mean RMSD for each cluster and finding the frame with the least difference from the mean. Further validation of the clustering was performed by creating a trajectory for each of the clusters containing all of the frames in each cluster and visually confirming the similarity within each cluster. Through visual analysis of the cluster representative frames, two clusters were determined to be highly similar and were thus combined. The unbound frames were then reintroduced as a single cluster resulting in a total of four clusters, observed to be top binding pose, side binding pose, bottom binding pose, and unbound (**Figure S8**). To get the conformation distribution of the apo DNA, the DNA structures from the DNA only simulations were clustered using Daura algorithm[71] with the 2.5 Å pair-wise RMSD cutoff of the DNA backbone. Three clusters were obtained, their representative structures are shown in **Figure S19** (left column). Conformational comparisons between the apo-form and ligand-bound structures of Pu24 G-quadruplex before and after DBD1 ligand-binding occurs are also shown in **Figure S19.**

Transition Path Theory and Markov State Model: Count matrices were then created for lagtimes (τ) of 1, 10, 20, 30 … 1000 ns by counting the number of observed transitions between discrete states such that the count of transitions from state *i* to state *j* (c_{ij}) is the sum of the number of times each of the trajectories were observed in state *i* at time *t* and in state *j* at time $t + \tau$, for all $t \leq t_{max} - \tau$ [72]. The count matrices were symmetrized (sym_{ij}) such that $sym_{ij} = sym_{ji} = \frac{c_{ij}+c_{ji}}{2}$ and then row-normalized ($norm_{ij}$) such that $norm_{ij} = \frac{sym_{ij}}{\sum_{j=1}^{j=n} sym_{ij}}$. For the purpose of determining the lag time at which the model has converged, the implied timescale of each binding state was calculated for all lagtimes and plotted (**Figure S9**). The implied timescale of the unbinding state is not included in the plot as the eigenvalue is always 1 and thus contributes no information [73]. Further validation that the model had been converged was performed through the Chapman-Kolmogorov test (**Figure S11**) [72]. The probability distribution (**Figure S12**) of the apo-form, side, top and bottom binding states as a function of increasing steps (250 ns/step) was also generated. The first data points show each state's initial probability distribution between 0 and 250 ns; the curvature shown by the subsequent datapoints indicate equilibrium was reached. This plot takes advantage of the Markov property (i.e. the memoryless component of the Markov model). The Markov property is given by the equation:

$$\lim_{k\to\infty}(P^k)_{i,j} = \pi_j$$

Where k is the step, P is the probability of the state i transitioning to state j and π_j is the jth value of a row vector π. The Markov property is satisfied assuming the probability of transitioning to future states is only dependent on the present state. By applying the Markov property to the transition matrix X containing P_{ij}, the resultant Markov model can forecast the probability distribution of each state at increasing step k until reaching equilibrium.

A network model (**Figure S13**) was then generated based on the count matrix at a lag time of 250 ns with the cutoff for a directed edge in the network being set at 300 transitions [74]. Thereafter, the mean first passage times (F_{if}) at a lag time of 250 ns and the standard deviations from lag time 250 ns to 750 ns were calculated according to the formula

$F_{if} = \tau + \sum_{j \neq f} P_{ij} F_{jf}$, with the boundary condition $F_{ff} = 0$, where τ is the lag time used to construct the transition matrix $P(\tau)$.

Order parameters to characterize DNA-drug binding pathway: Five ligand binding order parameters were calculated to characterize the DNA-binding process: hydrogen bond analysis, center-to-center distance (R), drug-base dihedral angle, ligand RMSD and MM-PBSA binding energy (ΔE). Hydrogen bonds are defined as the distance cutoff between hydrogen donor and hydrogen acceptor was set as 3.5 Å and donor-H-acceptor angle cutoff was set as 120º. The hydrogen bonds were calculated from the top/first, middle/second, bottom/third G-tetrad base layer and fourth triad layer. These base layers include the first (G4, G8, G13 and G17), second (G5, G9, G14 and G18), third (G6, G24, G15 and G19) G-tetrads and fourth (G20, A22 and G23) (**Figure 1**), where the three G-tetrads are referred to as the three layers with the 5' side representing the first layer of the G-quadruplex. Dihedral angle is defined as the dihedral angle between the plane of stable unbroken base-layers of the DNA that are close to drug binding site and the drug's benzofuran 2-ring plane. After aligning the DNA, the ligand RMSD was calculated with reference to the first frame of the trajectory. Center-to-Center distance (R) is defined as the length from the DNA center to the center of the drug molecule and the length between the two cations within the G-quadruplex structure. The distance between the K^+ ions present in the DNA G-quadruplex was defined as

K^+-K^+ distance. The MM-PBSA[75] (Molecular Mechanics Poisson Boltzmann Surface Area) module in the AMBER package (PB1 model with mBondi radii set, salt concentration of 0.15 M, and surface tension of 0.03780 kcal/Å^2) was used to analyze the energetics of the bound complexes to avoid the large energy fluctuation of explicit solvent[76]. The MM-PBSA binding energy for a system was calculated from three simulations: ligand only, DNA only and DNA-ligand complex using equation 1. It has four components in the equation 2: Gas phase Van der Waals interaction energy (VDW), Nonpolar solvation (SUR), electrostatic interaction (PBELE) and the change of the conformation energy for DNA and ligand. These terms were calculated using equation 3 and 4.

Eq 1: $\Delta E = E_{complex} - E_{DNA\_free} - E_{lig\_free}$

Eq 2: $\Delta E = \Delta E_{vdw} + \Delta E_{SUR} + \Delta E_{PBELE} + \Delta E_{comformation}$

Eq 3: $\Delta E_x = E_{x\_complex} - E_{x\_DNA\_complex} - E_{x\_lig\_complex}$, *x= vdw, sur and pbele*

Eq 4: $\Delta E_{Comformation} = E_{DNA\_complex} + E_{lig\_complex} - E_{DNA\_free} - E_{lig\_free}$

A recent study has shown that PB models make good prediction on the hydration free energy even for charged molecules when the relative solvation free energy is considered[77]. Systematic benchmarking studies up to 1864 crystal complexes have shown that relative MM-PBSA binding energy is a powerful tool to rank ligand binding affinity [78-82].

Structural fluctuation of the DNA and RNA: The root mean square fluctuations (RMSF) of each individual residue was calculated to characterize the local structural fluctuation of the apo form as well as that of the top, side, and bottom binding modes.

Alignment of most abundant clusters with experimental structures: The most abundant clusters were aligned in VMD using the RMSD calculator tool. The clusters were aligned with the solved NMR structures 2MGN and 5W77 and matching guanine residues within the G-tetrads within each G-quadruplex cluster.

Docking of DBD1 derivatives. Chemical derivatives of DBD1 ligand were designed in an attempt to enhance the binding of DBD1 to the top binding pose from our MD simulations. One combinatorial libraries of 625 derivative compounds were generated using the Interactive Enumeration Combinatorial Library tool in Maestro, with DBD1 acting as the core structure (**Figure S21**). In the library (**Figure S21**), substitution points S1 and S2 were assigned H, F, Cl, OMe or NMe_2 fragments and substitution points S3 and S4 were assigned H, F, Cl, NH_2 or CF_3 fragments, respectively. The generated compounds were then docked to the MD simulation top-binding site via the Glide XP Docking tool, using standard settings and procedures.

Results

Multiple binding modes were observed in free ligand binding simulations. Here, we probe a molecular basis for stabilizing the c-MYC G-quadruplex structure with a drug-like small molecule, DBD1. Starting from an unbound state at different locations near the G-quadruplex, we simulated thirty-three 1001 ns production runs for the G-quadruplex-ligand system. The convergence of the binding simulations was confirmed through our RMSD analysis (**Figures S4 and S5**), as described in the methods section, and a sampling plot was generated to trace the position of DBD1 throughout the length of all 33 trajectories (**Figure S3**). Additionally, the atom contacts between DBD1 and the Pu24 G-quadruplex were generated for the length of all 33 trajectories to further confirm the convergence of the simulations. The last snapshots of each simulated trajectory indicate the stability of the G-quadruplex structures because the G-tetrads were maintained (**Figure S7**). Of the 33 trajectories, the final binding poses at 1001ns were 23 top binding, 6 bottom binding, and 4 side binding.

Clustering analysis revealed 3 binding modes. Three binding modes (top intercalating, groove binding, and bottom stacking) were obtained through the clustering analysis described in the methods section (**Figure 2**). The first cluster, also the most abundant, is a top binding pose that consists of 55.9% of the simulation. This conformation is stable and is part of the top binding motif. DBD1, above the first G-tetrad layer (G4, G8, G13, G17), exhibits intercalation at the 5'-end of the G-quadruplex. The second cluster is a side binding pose that consists of 15.6% of the simulation and only exhibits minor intercalation between A21 and G23. The third cluster, consisting of 12.1% of the simulation, is a bottom binding pose. Altogether, they encompassed 83.6% of all the trajectories. The remainder of the simulation consisted of the unbound state (16.4%).

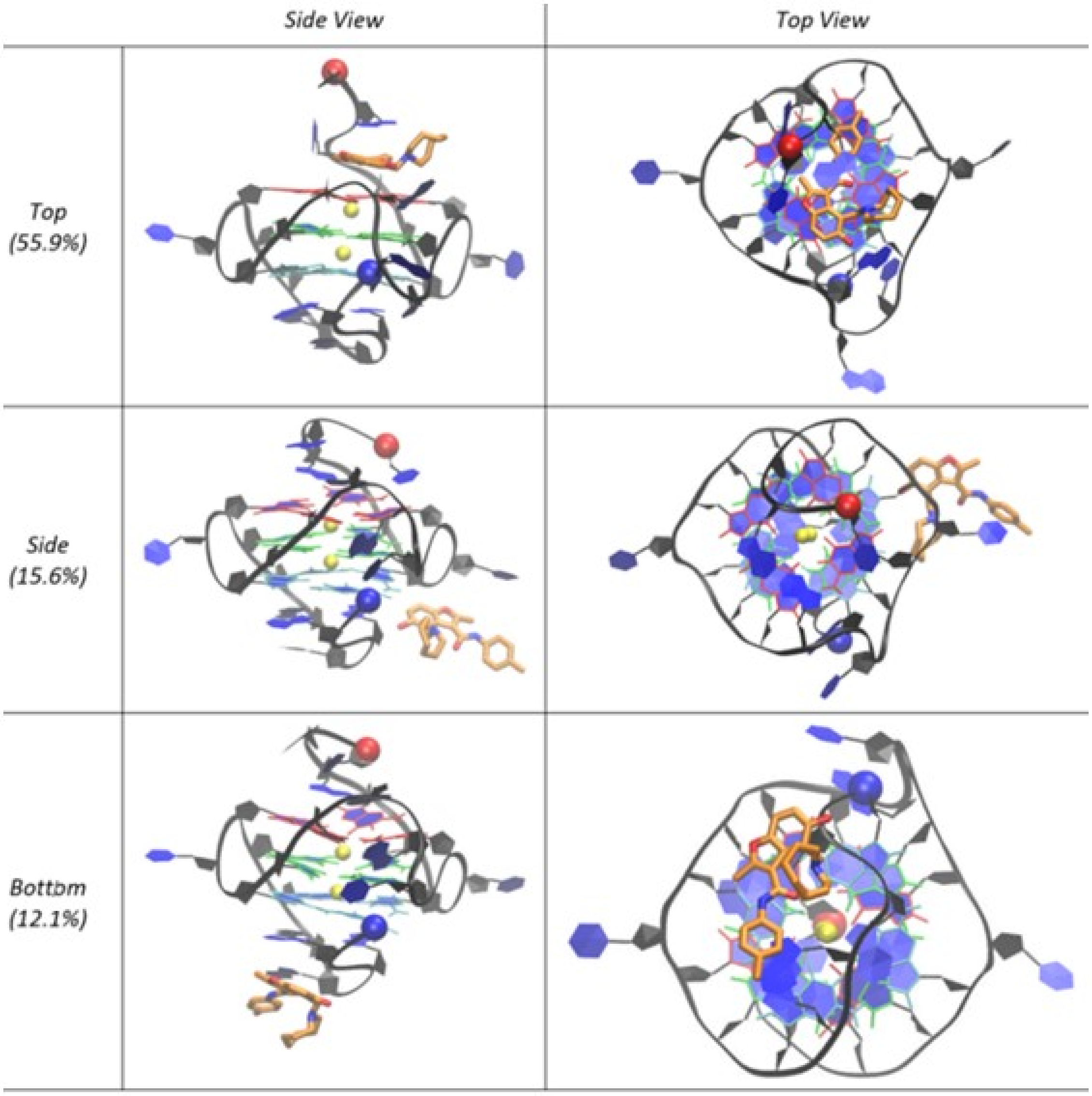


Figure 2: Representative complex structure of populated bound clusters from the K-means clustering analysis. Unbound cluster was not shown (16.4%). The three G-tetrad layers in the G-quadruplex are in red, green, and light blue for the top, middle and bottom layers, respectively. The ligand (DBD1) is in orange and K+ cations are represented by yellow balls.

Clustering analysis also revealed three clusters of the apo-form of Pu24 G-quadruplex. The six total representative clusters from the apo-form and bound Pu24 are represented in **Figure S19**. The most populous cluster of the apo-form of Pu24 is Cluster1 (A: 84.2 %), which shows good base pairing between nucleotide bases A3 and A12 (purple). After DBD1 binds to the bottom site of Pu24, the resulting cluster shows a significant decrease in population percentage (B: 12.1 %, a 72.0 % decrease). Additionally, the terminal nucleotide T1 (green) moves out of the plane of A3 and A12, in which the latter become slightly buckled. Consequently, the least populous cluster of the apo-form of Pu24 is Cluster 3 (E: 1.0 %), which shows base unpairing between A3 and A12. After DBD1 binds to the top site of Pu24, the resulting cluster shows a significant increase in population percentage (F: 55.9 %, a 58.9 % increase). Interestingly, the apo-form conformation population prior to side binding (C: 14.8 %) increases negligibly upon side-binding of DBD1 to Pu24 (D: 15.6 %, a 0.8 % increase). While poor base pairing between A3 and A12 is observed in the apo-form, tri-base pairing is seen between nucleotides T1-A3-A12. It seems that most of the population percentage transfer occurs from cluster 1 to Top binding cluster, with negligible transfer occurring from cluster 1 to Side-binding cluster. These data clearly suggest the ligand binding in this system follows the conformational selection theory.

Comparison of MD representative structures against 2MGN and 5W77 experimental structures shows significant similarities for the top binding mode and differences for the bottom binding mode. Comparisons between the two major binding modes (top and bottom) from the simulations and the two experimental complex structures are shown in **Figure 3-5**. Pu24 in complex with Phen-DC3 (PDB ID: 2MGN) was superimposed with the MD structure of Pu24 in complex with DBD1 in the top binding mode (**Figure 3**). Encouragingly, similar positioning of Phen-DC3 and DBD1 in the top G-tetrad was observed, suggesting a similar binding mechanism. Phen-DC3 remained relatively planar, whereas DBD1 exhibited an upward configuration of its nitrogen-containing azepane group towards the 5'-end loop of the experimental structure. Phen-DC interacts with all four residues of the first G-tetrad layer while DBD1 appears to mainly interact with only G8 and G13. Overall, we observed the G-quadruplex structures to be very similar when superimposed where minor G-quadruplex structural differences were observed at the 5’ end (T1G2A3) and the loop sequence (T10G11A12), of which base flipping of (5’-TGA) into the core of MD top was observed.

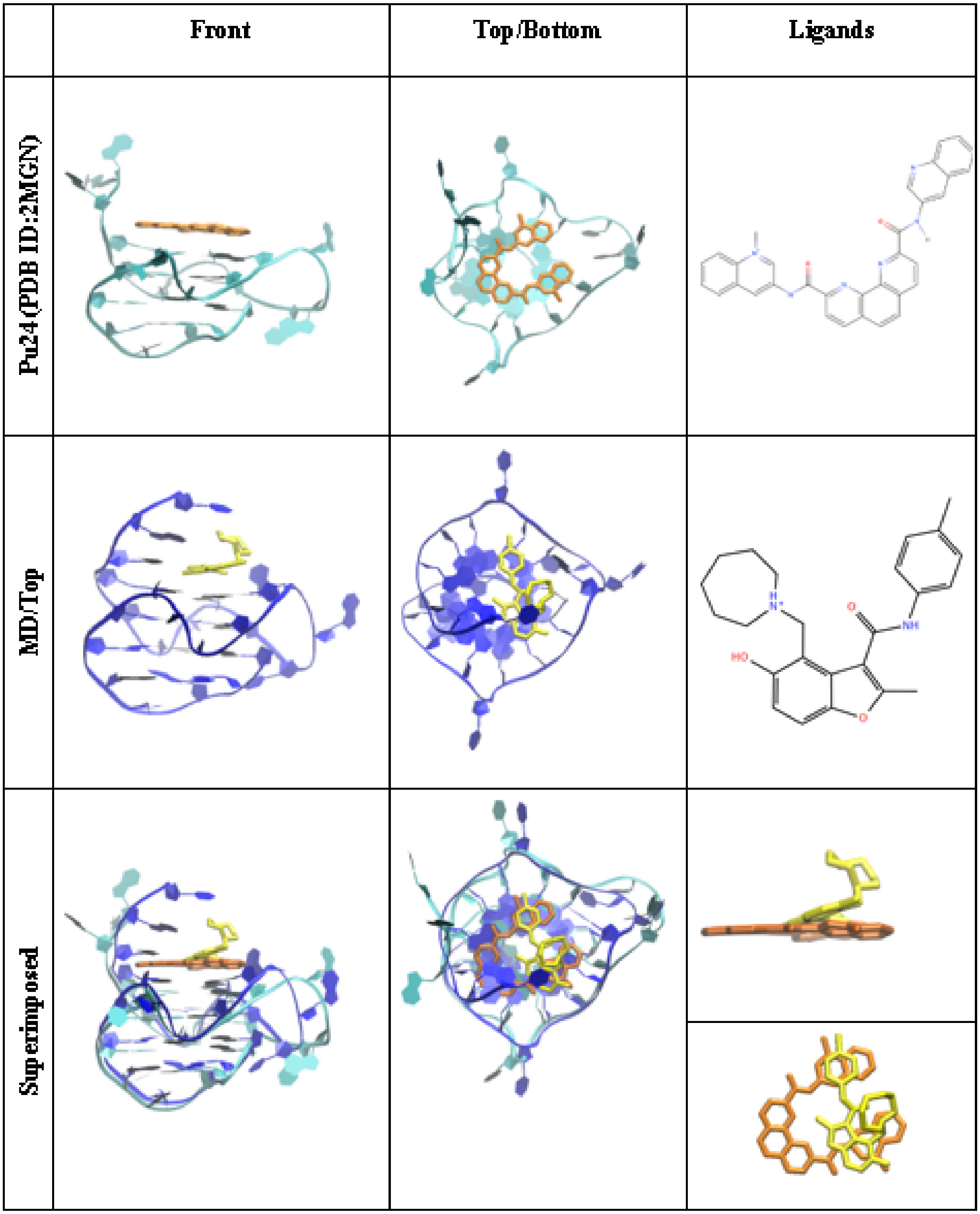


Figure 3: The top binding mode comparison between the experimental structure (PDB ID: 2MGN) and the MD Top structure of Pu24 in complex with a ligand. 2MGN is represented in cyan, MD Top is in blue respectively. Ligand Phen-DC3 in 2MGN is represented in orange and DBD1 in MD top is in yellow.

The Pu22 G-quadruplex in 5W77 has two molecules of DC-34 binding to the top and bottom positions. These two poses were superimposed with the top and bottom pose from our MD simulation of Pu24 in complex with DBD1 (**Figures 4 and 5, respectively**). Encouragingly, for the top mode, a very similar positioning of DC-34 and DBD1 was observed in the top G-tetrad (**Figure 4)**, suggesting a similar binding mechanism. For more detail, DBD1 showed its 1-benzothiapene group binding slightly deeper within its top binding pocket, losing planarity with DC-34 in 5W77. Additionally, the terminal methyl group of DBD1 projects rightwards of the trifluoromethyl group of DC-34, nearly stacking with the azepane group of DC-34. Due to this orientation, DC34 interacts with G4 and G13 while DBD1 interacts with G8 and G13. Structural deviations in the 5'-TGA sequence between Pu22 in 5W77 and MD Pu24 are not as significant as is seen between Pu24 in 2MGN and MD Pu24 in **Figure 3**. These differences are greatest in their 3'-end loops ($A^{21}A^{22}$-3') and a short sequence ($A^{18}A^{19}$), of which base flipping is not apparent here.

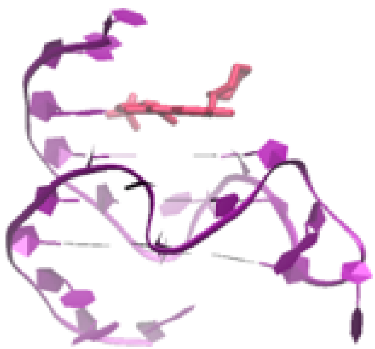
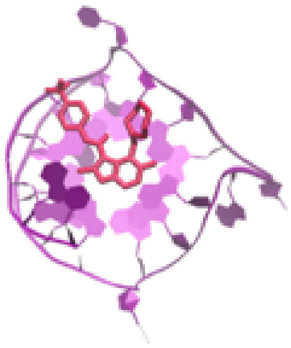
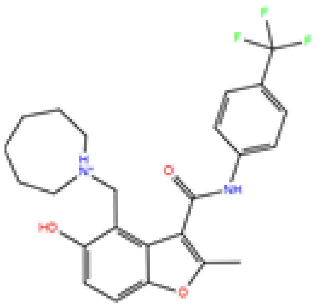
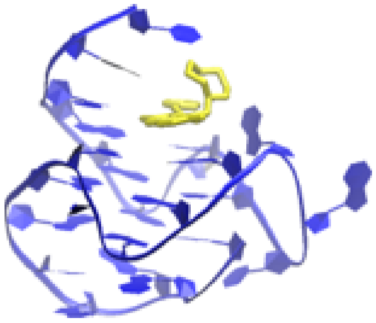
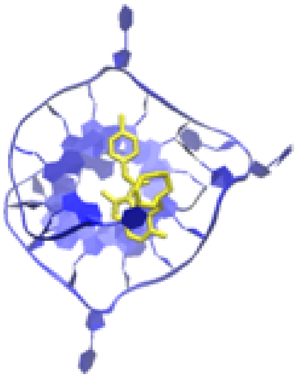
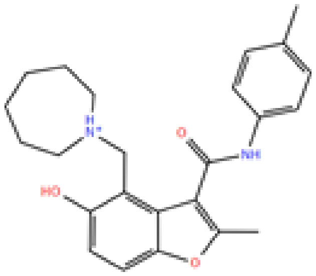
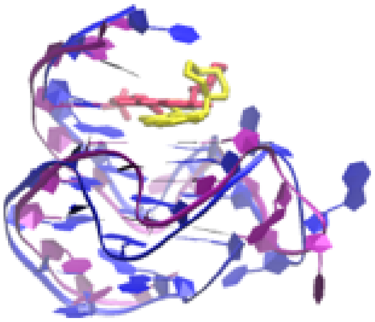
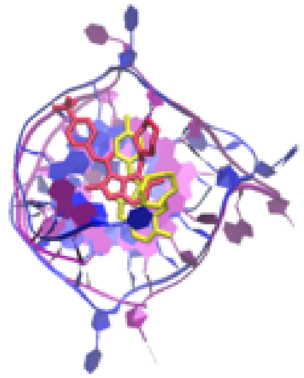
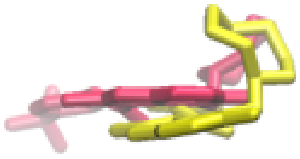
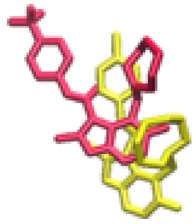
	Front	Top/Bottom	Bottom
Pu22 (PDB ID : 5W77)			
MD Top			
Superimposed			 

Figure 4: The comparison between the experimental structure (PDB ID: 5W77) and the MD Top structure. While the DNA sequence in 5W77 is Pu22, the DNA sequence in MD Top is Pu24. Ligand DC-34 in 5W77 is represented in purple and ligand DBD1 in MD Top is in yellow.

However, very different positioning of DC-34 and DBD1 in the bottom mode was observed in **Figure 5**. The ligand DC-34 binds deeper into the bottom-binding pocket of than DBD1; the shortened 3’-loop of Pu22 G-quadruplex allows bottom binding between the third G-tetrad and the 3’ terminal segment, whereas the 3’-loop end of the MD Pu24 structure connects to the third G-tetrad layer and restricts bottom-binding of DBD1 between third G-tetrad and the 3’ terminal segment. Because DC-34 and DBD1 share the same molecular scaffold, the difference in bottom binding modes are due to the structural difference of the 3’ region between Pu22 and Pu24, which will be discussed in greater detail later. Another structure difference between Pu22 in 5W77 and our MD simulation of Pu24 in the bottom mode was in the connecting loop ($T^{10}G^{11}A^{12}$), where tyrosine 10 and guanine 11 experience base-stacking relative to each other in the MD Pu24/bottom.

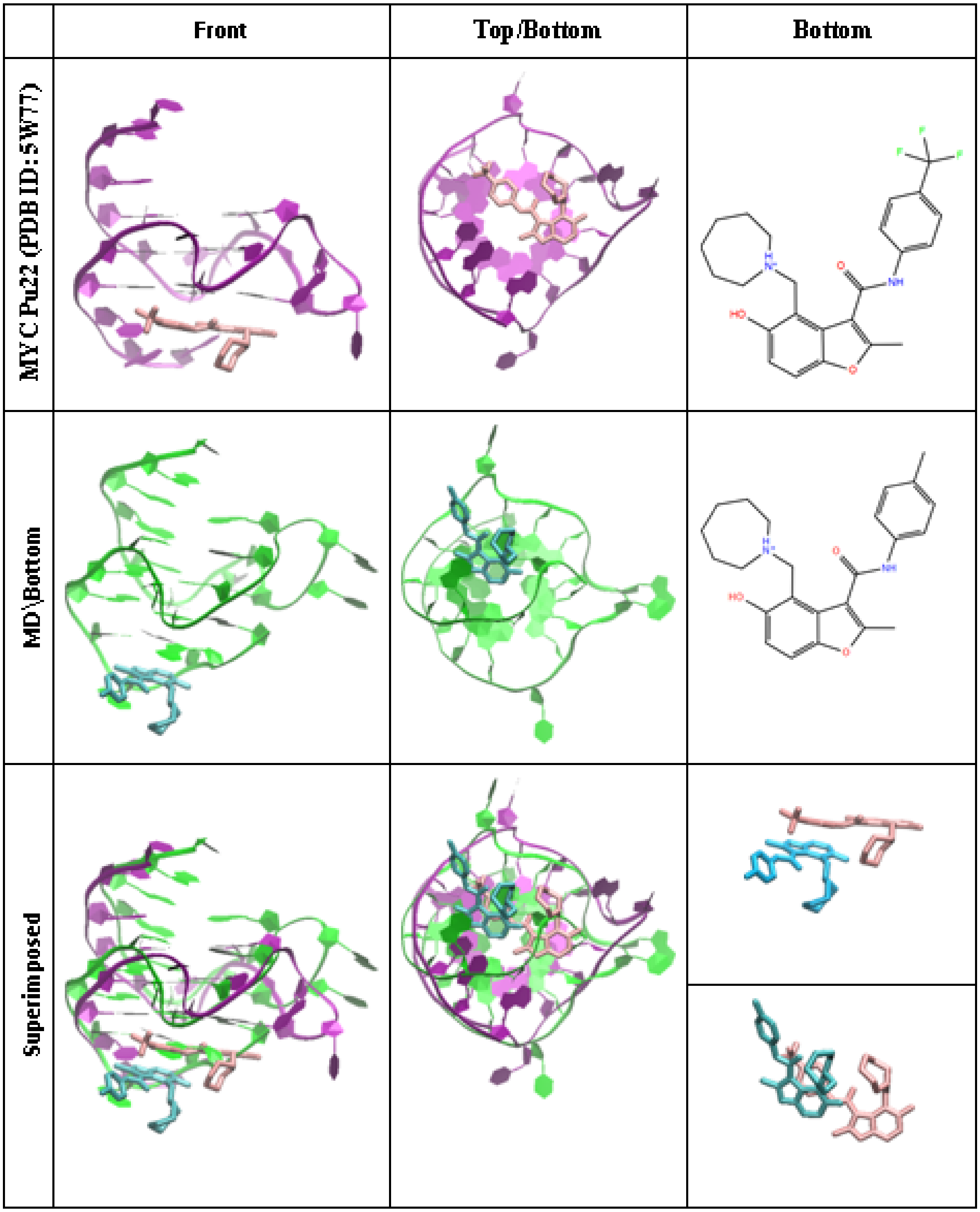


Figure 5: The bottom ligand binding mode comparison between experimental structure ((PDB ID: 5W77) and the MD Bottom structure. While the DNA sequence in 5W77 is Pu22, the DNA sequence in MD Top is Pu24. Ligand DC-34 in 5W77 is represented in pink and ligand DBD1 in MD Top is in cyan.

MM-PBSA binding energy calculations determined that the top binding mode was the most energetically favorable. MM-PBSA binding energy calculations were conducted for the three ligand binding modes in order to determine the relative stability of the three major binding modes (**Table 2**).

Table 2. MM-PBSA binding energies (kcal/mol) of DBD1 in the Top, Bottom, and Side binding modes.

Position	ΔE_{VDW}[a]	ΔE_{SUR}[b]	ΔE_{PBELE}[c]	ΔE_{PBTOT}[d]	ΔConf[e]	ΔE_{TOT}[f]	$\Delta\Delta E_{TOT}$[g]
Top	-52.0±0.3	20.3±0.3	5.9±0.4	-25.9±0.4	-16.9±4.6	-42.9±4.5	0
Bottom	-14.7±1.0	8.3± 1.2	1.9±1.0	-4.5.±0.8	-12.4±1.9	-16.9±1.8	26.0
Side	-14.3±0.5	8.4±1.9	1.4±0.2	-4.5±0.5	-12.4±2.7	-16.9±2.6	26.0

[a] Gas phase van der Waals energy (VDW)
[b] Nonpolar solvation (SUR=PBSUR+PBDIS)
[c] Solvation and gas phase electrostatic energy (PBELE=PBCAL + ELE)
[d] PB Solvation and gas phase energy (PBTOT=VDW+SUR+PBELE)
[e] Conformation energy change upon complex formation (Conf)
[f] Total binding energy in water (PBTOT + Conf)
[g] Relative binding energy

The binding energy calculations indicated that the most energetically favorable binding pose was the top binding mode (-42.9±4.5 kcal/mol) followed by both the bottom (-16.9±1.8 kcal/mol) and side (-16.9±2.6 kcal/mol) binding modes which had similar binding energies. Van der Waals forces play a major role in the stability of the binding of DBD1 to Pu24 as can be seen when comparing the van der Waals forces of the top binding mode (-52.0±0.3 kcal/mol) to that of the bottom (-14.7±1.0 kcal/mol) or side (-14.3±0.5 kcal/mol) binding modes. The change in binding energy ($\Delta\Delta E_{TOT}$) between the top binding mode and the other two binding modes is less than the difference in van der Waals energy, indicating that the van der Waals interactions make up the majority of the total MM-PBSA binding energy for all three binding poses. We see that the difference in PB_{TOT}, PB solvation and gas phase energy, also plays a lesser but non-negligible role in the difference between the binding energies of top (-25.9±0.4 kcal/mol), bottom (-4.5±0.8 kcal/mol), and side (-4.5±0.5 kcal/mol) binding modes.

Multiple pathways were observed in free ligand binding simulations. The 33 simulated trajectories can be further classified into different binding pathways. First, there are 16 trajectories that exhibit the transition shown from unbound directly to the top binding state (**Figure S14**). Second, 9 trajectories show the ligand going from the unbound state to the side transition state to the final top binding state (**Figure

S15). Third, 3 trajectories indicate a transition from either top or bottom to the side binding state (**Figure S16**). Fourth, there are 2 trajectories that indicate the transition from unbound to bottom which combined with the aforementioned transitions shows the possibility of an unbound to bottom to side to top transition if the trajectories were to be extended (**Figure S17**). Fifth, some of the reverse pathways can be observed such as 3 trajectories exhibiting the transition from side back to bottom binding (**Figure S18**). Clearly, these observed pathways support our MSM (**Figure 6**).

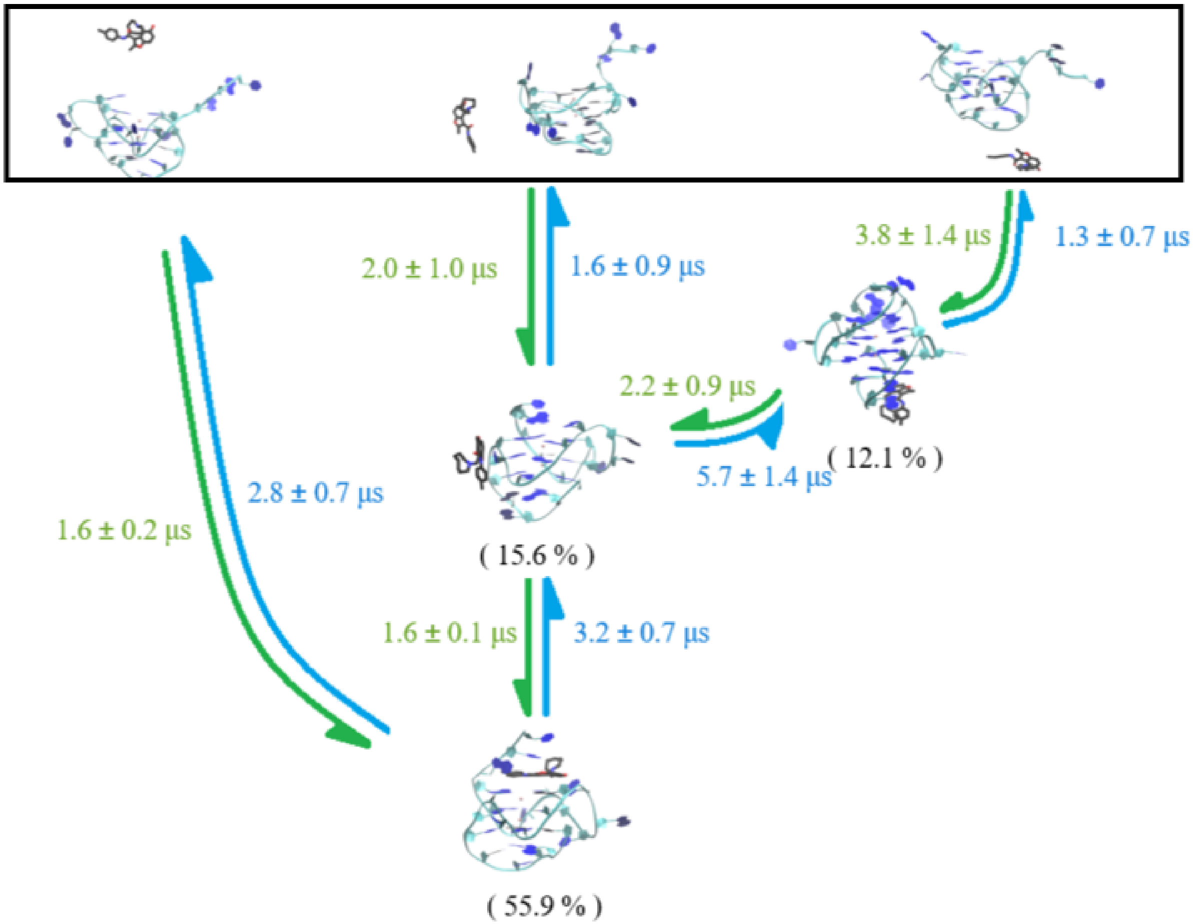


Figure 6. MSM of DBD1 binding to Pu24 G-quadruplex. The top row consists of representative structures of the unbound state. The middle row consists of the two intermediate states, side and bottom. The bottom row consists of DBD1 binding to the top site of the G-quadruplex. The mean first passage times between the four states (unbound, bottom, side, and top) are annotated in the same color as the arrow directing the transition. DBD1 and the Pu24 G-quadruplex are colored black and blue/cyan respectively.

MSM analysis reveals three parallel binding pathways from unbounded state to the top binding mode, the most thermodynamically stable state. The clustering identified four macrostates (unbound, top, side and bottom binding) and MSM analysis was performed on those states using transition path theory, as mentioned in the methods section, to obtain binding pathway information. Identification and verification of the optimal lag time were performed using the implied timescales and Chapman-

Kolmogorov test, also discussed in the methods section (**Figures S10 and S11**). A network model with the optimal lag time (250 ns) was presented with the transition counts (**Figure S13**): the approximate ratios of the interstate fluxes were 1:3 for unbound to top binding, 4:3 for unbound to side binding, 1:1 for unbound to bottom binding, 1:3 for side binding to top binding, and 1:2 for bottom binding to side binding. To simplify interpretation, the mean first passage time between each of the two connected states, connections being defined as any two states that had at least 1 transition, was calculated. The transitioning of the other states towards the top binding mode, the most thermodynamically stable state, was analyzed and presented in a reorganized MSM of DBD1 binding to the Pu24 G-quadruplex (**Figure 6)**. Figure 6 shows organization of the states from top to bottom begins with unbound on top and then from least abundant to most abundant (abundance is displayed in parentheses). Overall transition times for each given path are organized from fastest (left) to slowest (right). It can be clearly seen that the transition from the unbound state directly to the final binding pose is the fastest while transitions involving transition states are significantly slower. Transition to the bottom transition pose requires transition to the side transition pose in order to reach the final binding pose. Interestingly, three distinct binding pathways to the most stable binding mode (top binding) were obtained. The first major binding pathway was the direct transition from unbound to the final top binding state (60% of flux). The second and third binding pathways involve additional transitions through the side binding transition state (27% of flux) and both the bottom binding transition state and the side binding transition state (13% of flux) respectively. We can see from our results that the ligand has multiple pathways, some more favored than others, to reach the final binding pose and these pathways are observable in the original trajectories, further supporting our results.

A representative trajectory for three key pathways was chosen for detailed characterization using order parameters. A representative trajectory for the first three pathways was chosen for further characterization using some order parameters (**Figure 7-9**). We measured hydrogen bonds, center-to-center distance (D), drug-base dihedral angle, receptor and ligand RMSD and MM-PBSA binding energy

(ΔE) as descripted in the method section. The top stacking mode was the most energetically favorable and stable structure according to the RMSD and MM-PBSA binding energy.

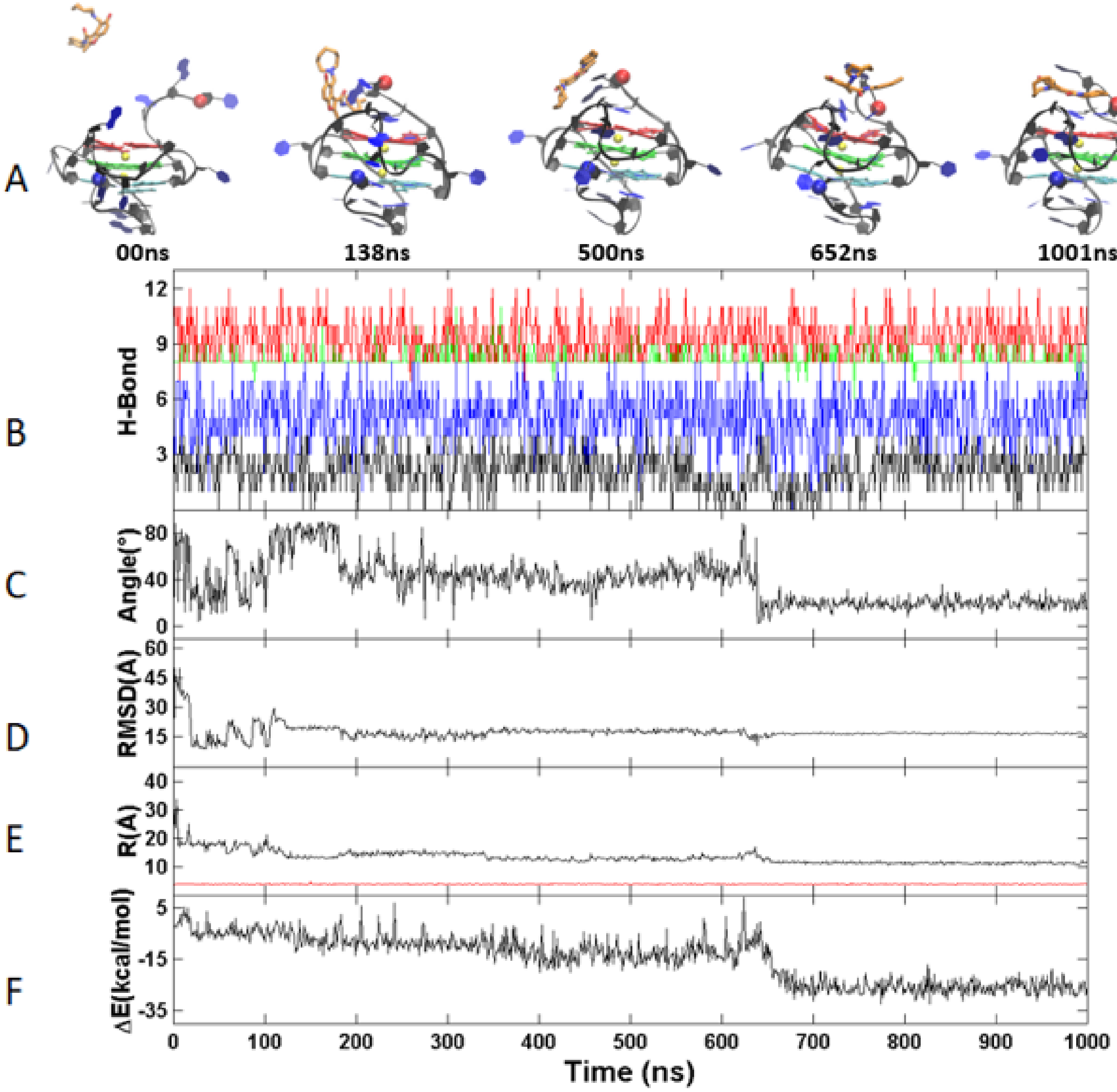


Figure 7. Order parameters calculated from a representative trajectory of the primary binding pathway of DBD1 to the top position of Pu24 G-quadruplex. Top-bottom: **(A)** Representative structures with time annotation. 5' and 3' are indicated by a red and blue ball, respectively. K^+ ions are represented in yellow. **(B)** Hydrogen bonds in the first (red), second (green), third (blue) G-tetrad and fourth (black) of G-triad layer of quadruplex (H-bond), **(C)** drug-base dihedral angle, **(D)** ligand RMSD, **(E)** center-to-center distance (R/black) and K^+-K^+ distance (R/red) and **(F)** MM-PBSA binding energy (ΔE).

Figure 7 shows unbound DBD1 nearly reaching the top binding position at 138ns and transitioning fully into this position by 652ns. Hydrogen bonding remains relatively stable throughout the simulation with about 10, 9, 6, and 3 hydrogen bonds in the first, second, third G-tetrad, and fourth triad layers of the G-quadruplex, respectively, suggesting little change in the G-quadruplex scaffold. The drug-base dihedral begins at ~80 degrees, decreases to ~40 degrees at 200ns, and stabilizes at 20 degrees at 650ns and throughout the remaining simulation. This highlights DBD1's intercalation between the top G-tetrad layer and the 5'-end loop. The DBD1 ligand RMSD stabilizes at ~20 Å by 150ns and remains stable throughout the simulation. Center-to-center potassium ion distancing between the ligand and the G-quadruplex stabilizes by 100ns and the potassium ions distances remains stable throughout the simulation. MMPBSA binding energy stabilizes at ~(-25 kcal/mol) by 650ns after DBD1 reaches the top binding site.

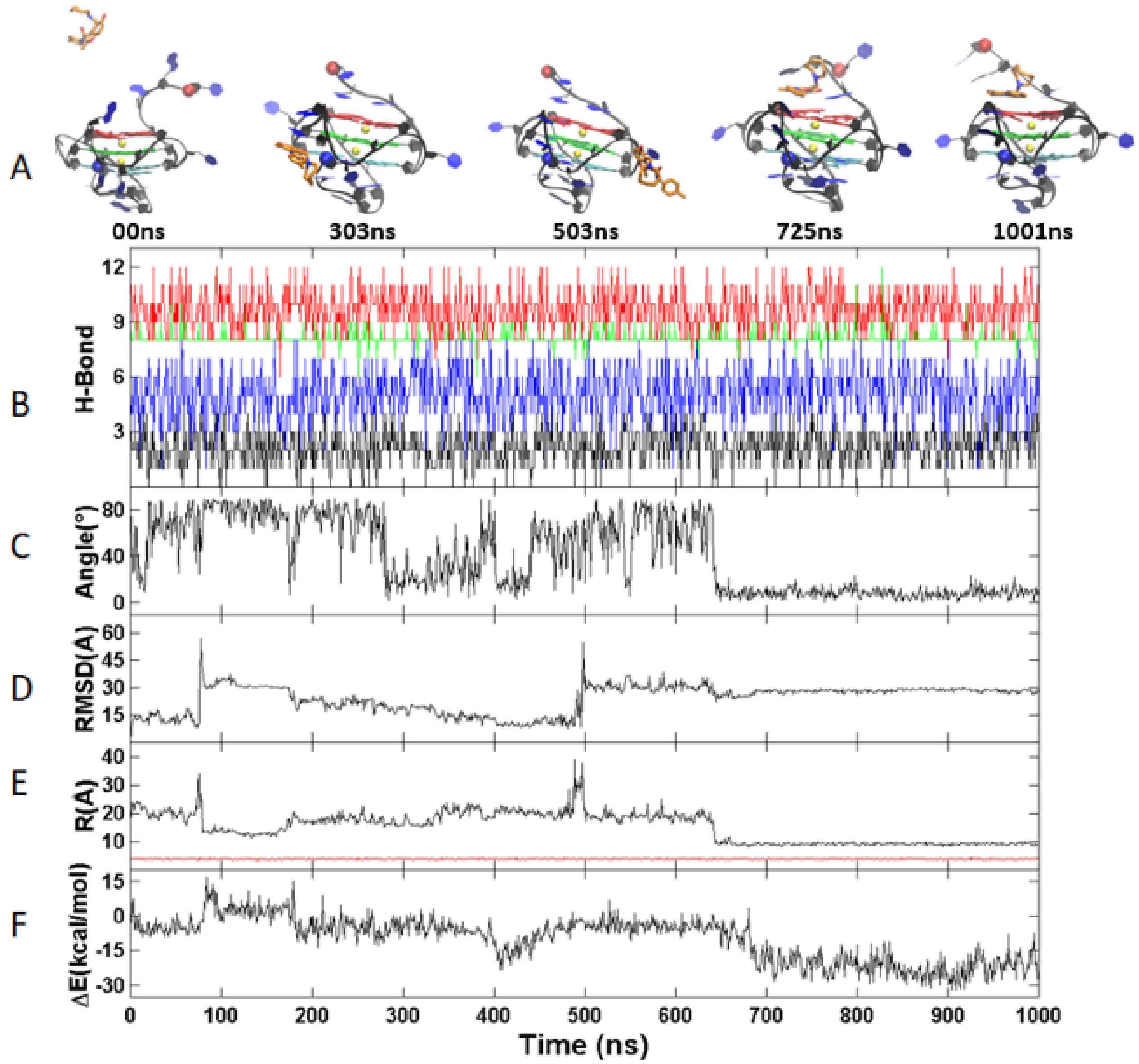


Figure 8. Order parameters calculated from a representative trajectory of the primary binding pathway of DBD1 to the top position of Pu24 G-quadruplex. Top-bottom: **(A)** Representative structures with time annotation. 5' and 3' are indicated by a red and blue ball, respectively. K^+ ions are represented in yellow. **(B)** Hydrogen bonds in the first (red), second (green), third (blue) G-tetrad and fourth (black) of G-triad layer of quadruplex (H-bond), **(C)** drug-base dihedral angle, **(D)** ligand RMSD, **(E)** center-to-center distance (R/black) and K^+-K^+ distance (R/red) and **(F)** MM-PBSA binding energy (ΔE).

Figure 8 shows unbound DBD1 that transitions to the side-binding site by 303ns and to another side-binding site at 503ns before transitioning to the top-binding position at 725ns. The hydrogen binding analysis here shows little change in the G-quadruplex scaffold, as in **Figure 7**. The dihedral angle initially averages 80 degrees, decreases to ~20 degrees at 300ns, increases again to ~80 degrees between 300ns and 650ns and stabilizes at 10 degrees at 650nsand throughout the remaining simulation. The DBD1 RMSD begins at 15 Å, increases sharply to ~30 Å at 100ns, decreases slowly until 500ns before spiking to ~30 Å again and then stabilizes with less fluctuation at 650ns until the end of the simulation. Center-to-center distance stabilizes by 650ns, while the distance between the two potassium ions remains stable for the entire simulation. MMPBSA energy stabilizes at ~(-25 kcal/mol) by 650ns where the ligand has reached the stable top binding position.

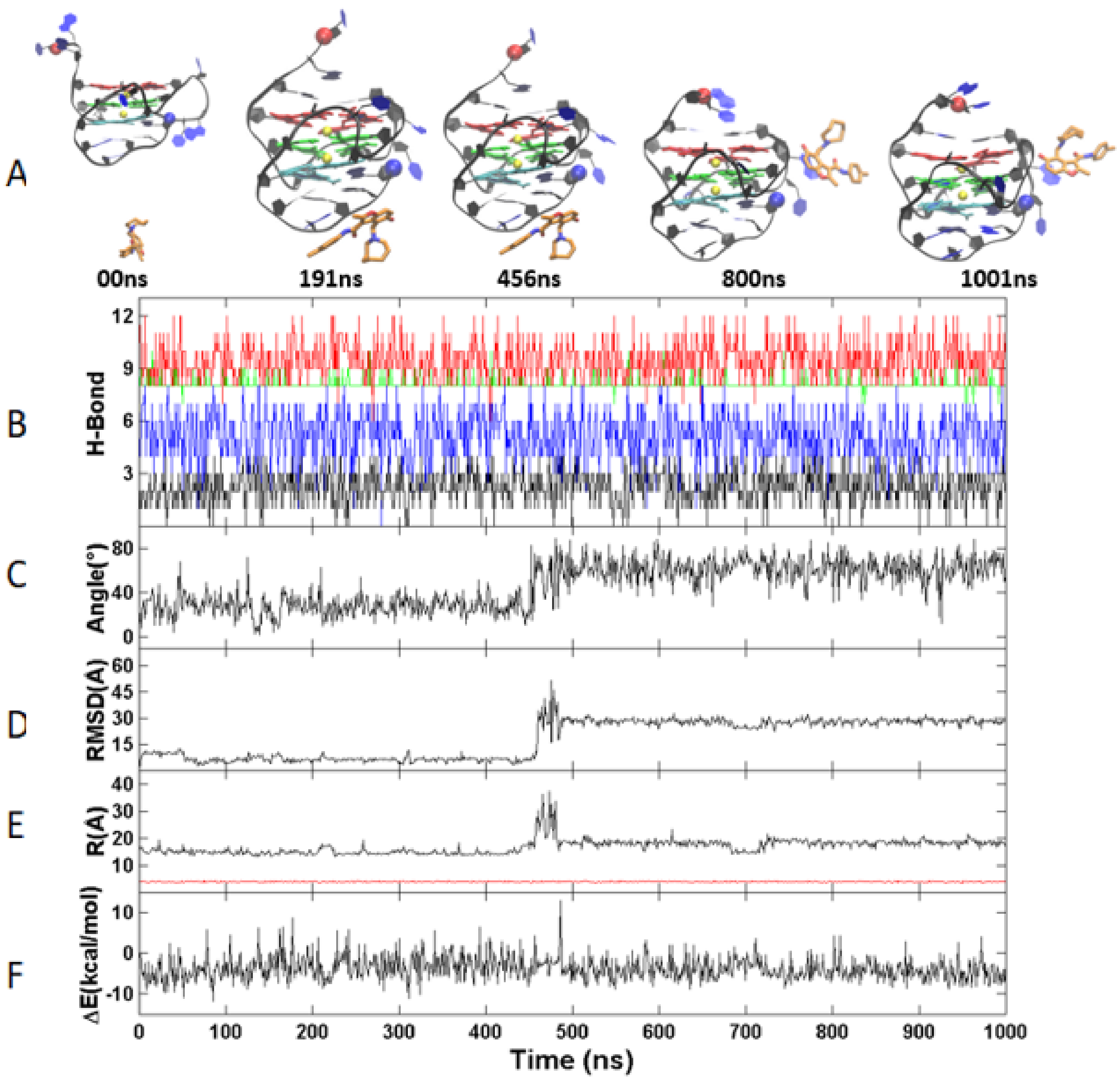


Figure 9. Order parameters calculated from a representative trajectory of the primary binding pathway of DBD1 to the side-binding position of Pu24 G-quadruplex. Top-bottom: **(A)** Representative structures with time annotation. 5' and 3' are indicated by a red and blue ball, respectively. K^+ ions are represented in yellow. **(B)** Hydrogen bonds in the first (red), second (green), third (blue) G-tetrad and fourth (black) of G-triad layer of quadruplex (H-bond), **(C)** drug-base dihedral angle, **(D)** ligand RMSD, **(E)** center-to-center distance (R/black) and K^+-K^+ distance (R/red) and **(F)** MM-PBSA binding energy (ΔE).

Figure 9 shows unbound DBD1 before it reaches the bottom-binding position at 191 ns and transitioning to the side-binding position at 800 ns. The hydrogen binding analysis here shows little change in the G-quadruplex scaffold, as seen in **Figures 7 and 8**. The dihedral angle initially averages ~30 degrees, increasing and stabilizing to ~70 degrees at 450ns throughout the remaining simulation. The DBD1 RMSD starts at 15 Å before increasing sharply to 30 Å at 450ns and stabilizes for the remaining simulation. Center-to-center distance is stable throughout the entire simulation except for a small fluctuation at ~450ns, while the distance between the two potassium ions remains stable throughout the simulation. MMPBSA energy fluctuates between 0 kcal/mol and -10 kcal/mol for the entire simulation.

Trends in **Figures 7-9** showed a lack of change in H-bonding within the G-quadruplex structure scaffolding; the G-quadruplex structure conformations were not significantly altered by ligand-binding. The center-to-center potassium ion distancing in the G-quadruplexes stays relatively small in all three simulations, suggesting that the core structures do not undergo significant conformational change. The nucleotide sequence ($T^{10}G^{11}A^{12}$) underwent noticeable conformational changes in all three systems, where base flipping was apparent in the ($T^{10}G^{11}A^{12}$) sequence. The top-binding position in all three trajectories has a lower dihedral angle and binding energy after DBD1 intercalates between the 5'-loop and the top G-tetrad layer; this suggests that DBD1 has geometric-favorability within the top-binding site, resting just above the first G-tetrad layer. The top binding position exhibits planarity with the G-quadruplex scaffold, which is not seen in either the bottom- or side-binding positions. This planarity allows DBD1 to intercalate here, serving an important factor in determining the most stable binding position. Our findings suggest that the most favorable binding position is top-stacking while the intermediate positions, side and bottom ultimately transition to the top-binding position of the G-quadruplex.

Geometric characterization of the three core G-quartets showed little change upon ligand binding. Literature suggests that the overall helical structure is most accurately defined by the H-twist and H-rise parameters. We compared these parameters in the top three major binding modes of DBD1 binding to the DNA G-quadruplex to the apo form of the DNA G-quadruplex system. Our observation of the DNA G-quadruplex core shows that each G-quartet has a right-handed helical rotation (**Figure 10A-B**). As detailed in the methods section, the rise, H-rise, and H-twist of the apo form DNA G-quadruplex and for each major binding mode (top, bottom, and side) was calculated (**Table 3**).

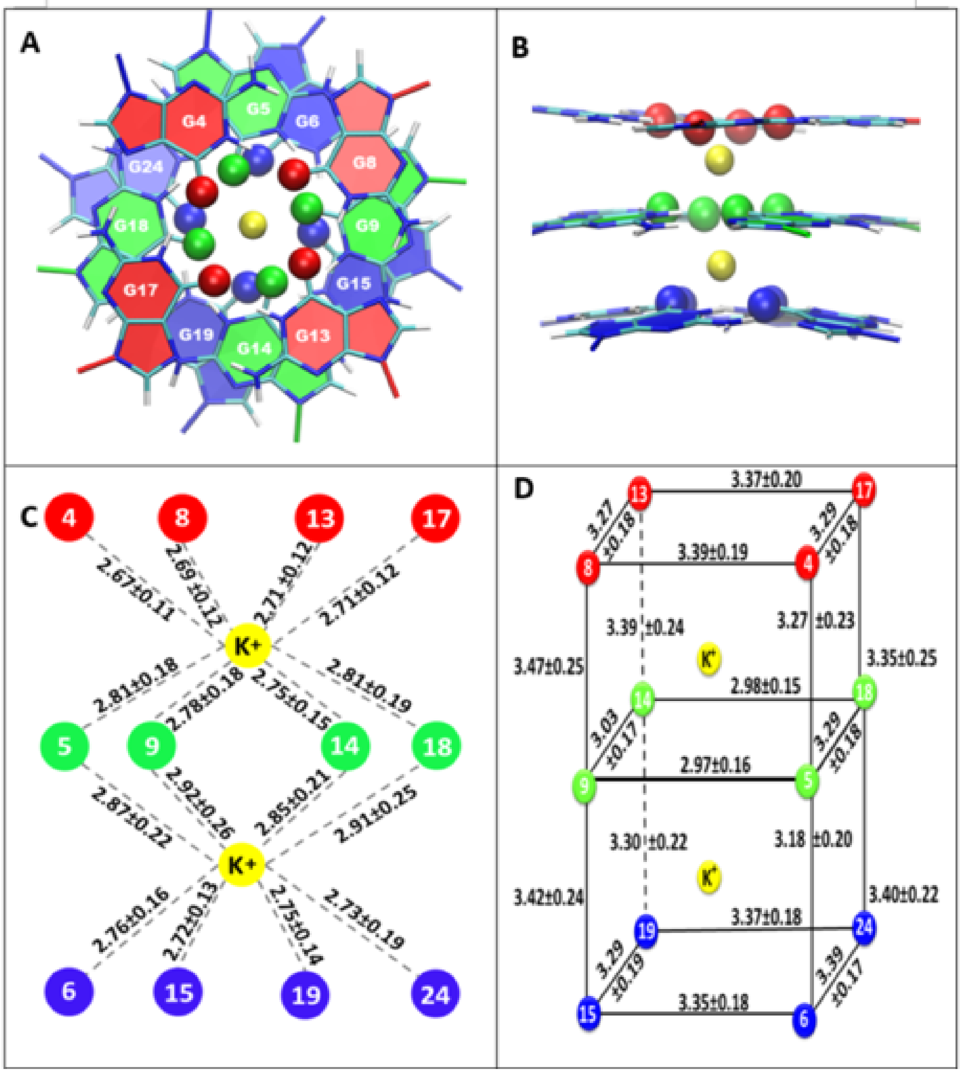


Figure 10. Oxygen-Potassium Distance Parameters in the 3-layers of the DNA G-quadruplex for the Pu24 apo form: Top view of the three G4 layers (**A**): Oxygen is represented by a colored ball and K^+ cations are represented by a yellow ball; Side view of the three G4 layers (**B**); Distance (Å) of the oxygen from each residue to the nearest K^+ cations(**C**); Distance (Å) of each oxygen relative to the nearest side (**D**).

Table 3. Pu24 G-quadruplex G4 layer geometry parameters. 1-3: top, middle and bottom G4 layer, respectively.

Layers	Parameter	Apo Form	Top Binding	Bottom Binding	Side Binding
3:2	Rise[1]	3.48±0.08	3.47±0.07	3.48±0.08	3.47±0.07
3:2	H-Rise[1]	3.46±0.08	3.45±0.07	3.46±0.08	3.46±0.07
3:2	H-Twist[2]	24.05±2.47	24.32±2.60	23.72±2.70	24.18±2.60
2:1	Rise[1]	3.42±0.08	3.43±0.08	3.44±0.08	3.43±0.07
2:1	H-Rise[1]	3.40±0.08	3.41±0.08	3.42±0.08	3.41±0.08
2:1	H-Twist[2]	26.29±3.04	25.94±2.98	26.12±3.13	25.91±3.10

[1] Distance measured in Å

[2] Angle measured in degree

The rise, H-rise and H-twist averaged over the two-layer steps (i.e. bottom to middle and middle to top) of the apo form are ~3.45 Å, ~3.43 Å and ~25.17˚, respectively. It was clear that while the average H-rise of the DNA G-quadruplex is slightly larger than that in standard B-DNA (3.32 Å) by 0.11Å, the H-twist of the DNA G-quadruplex is slightly smaller than that in standard B-DNA (34.3˚) by 9.1 ˚. It makes sense that the smaller the H-twist, the larger the H-rise. When comparing the two layer steps, the rise (~3.48 Å) and H-rise (~3.36 Å) of layers 3:2 are slightly larger than those (~3.42 Å and ~3.40 Å) of layers 2:1 by ~0.06 Å, the H-twist (~24.05˚) of layers 3:2 is slightly smaller than that (26.29˚) of layers 2:1 by ~2 ˚. These slight differences might be caused by the environment of the bottom layer (layer 3), which is different from the environment of the top layer (layer 1). The binding of DBD1 to the DNA G-quadruplex causes little changes to these parameters in each of the major binding modes when compared to the apo form: top (~3.45 Å, ~3.43 Å and ~25.13˚), bottom (~3.46 Å, ~3.44 Å and ~24.92), and side (~3.45 Å, ~3.43 Å and ~25.04˚). From comparing the values for H-Twist between layers, we observed a slightly smaller rotation required to align the middle quartet to the top quartet compared to the rotation required for the alignment of the bottom quartet to the middle quartet, which is a trend that did not change much upon ligand binding. Thus, the closely comparable averages in each system provide qualitative support that the helical structure of the DNA G-quadruplex was maintained throughout the simulations.

The interaction between K+ and the G-quadruplex did not change with ligand binding. There are many factors that contribute to the overall stability of the DNA G-quadruplex structure including the coordination of O6 carbonyls by cations from within the ion pore. Literature reports that after the bipyramidal antiprism is formed as a result of cation stabilization, the O6 carbonyls maintain an average inter-quartet distance of 3.3 Å[83]. This work defined some parameters to understand the position of the potassium cation relative to the surrounding G4 DNA (Figure 10; Table S2-S3). The distance between both potassium ions within the G-quadruplex ion pore and the eight neighbouring oxygen atoms (**Figure 10C**) were calculated (**Table S2**). The apo form of the DNA G-quadruplex had an average oxygen-potassium distance of 2.67 Å, 2.83 Å, and 2.72 Å, for the top, middle, and bottom G-quartets, respectively. For both calculations, the similar mean and low standard deviations provide support that our simulation parameters were appropriately set. Compared to the apo form, the major binding modes of DBD1 in complex with the DNA G-quadruplex showed no significant change in potassium-oxygen distances. There averages of each distance calculated for the top, bottom, and side binding poses were 2.76 Å, 2.77 Å, 2.77 Å, respectively, and 2.76 Å for the apo form. From this it is clear there is little variation between modes or upon DBD1 binding.

The distance between each G-quartet oxygen relative to one-another was also calculated for both the apo form (**Figure 10D**) and complex systems of the major binding modes (**Table S3**). Analysing the calculation of the distance between each oxygen residue lining the ion pore of the G-quadruplex core showed no significant difference in oxygen-oxygen distance upon ligand binding. For the apo form, the average distance between each residue of the same G-tetrad was 3.24 Å, and the average distance between each residue to the residue of the neighbouring G-tetrad was 3.32 Å. For the major binding modes, the average distance between each residue of the same G-tetrad was 3.24 Å, 3.27 Å and 3.26 Å, for the top, bottom, and side biding mode, respectively. The average distance between each residue to the residue of the neighbouring G-quartet was 3.37 Å, 3.35 Å and 3.36 Å, respectively, for the top, bottom, and side binding modes. These calculations clearly show that DBD1 binding did not significantly affect the interaction between the potassium cations and the G-quadruplex.

RMSF analysis indicated that all three of the binding poses of the structure complex of the c-MYC G-quadruplex with DBD1 had slightly greater fluctuation than that of the apo form; binding of DBD1 greatly increased fluctuation in residues 10 to 12. RMSF calculations were performed to identify fluctuation by residue across all 33 complex trajectories and all 5 apo form trajectories. The RMSF plot (**Figure 11**) revealed that the apo form had less fluctuation overall compared to any of the three complexed states and that the average fluctuation in decreasing order is top, side, bottom, and apo form. Two noticeable peaks are evident in the residues that comprise loop 1 (T1-A3) and loop 2 (T10-A12), while loop 3 (A20-G23) shows minimal fluctuation. The apo form showed a ~5.8 Å peak fluctuation for loop 1 and a ~2.7 Å peak fluctuation on residue T10 for loop 2. The complexes for top, side, and bottom binding states showed peak fluctuations of ~6.9, ~6.9, ~6.0 Å respectively for loop 1. For loop 2, peak fluctuations were observed at ~4.5, ~4.4 and ~4.2 Å at residue G11, A12, and A12 respectively for the top, side, and bottom binding complexes. Between the three complexed structures, we observe that the loop 2 peak fluctuation residues differ between the top binding state and the other two binding states. Loop 2 is closely located to the binding site of DBD1 in the top binding position and this may play a role in why the fluctuation is greater for the top binding state. Additionally, the fluctuations in loop 1 and 2 of the apo form are notably lower than that of the complexes when compared to the rest of the RMSF values. As the overall RMSF of the apo form is lower than that of the complexed structures, this suggests that the binding of DBD1 induces significant conformational change.

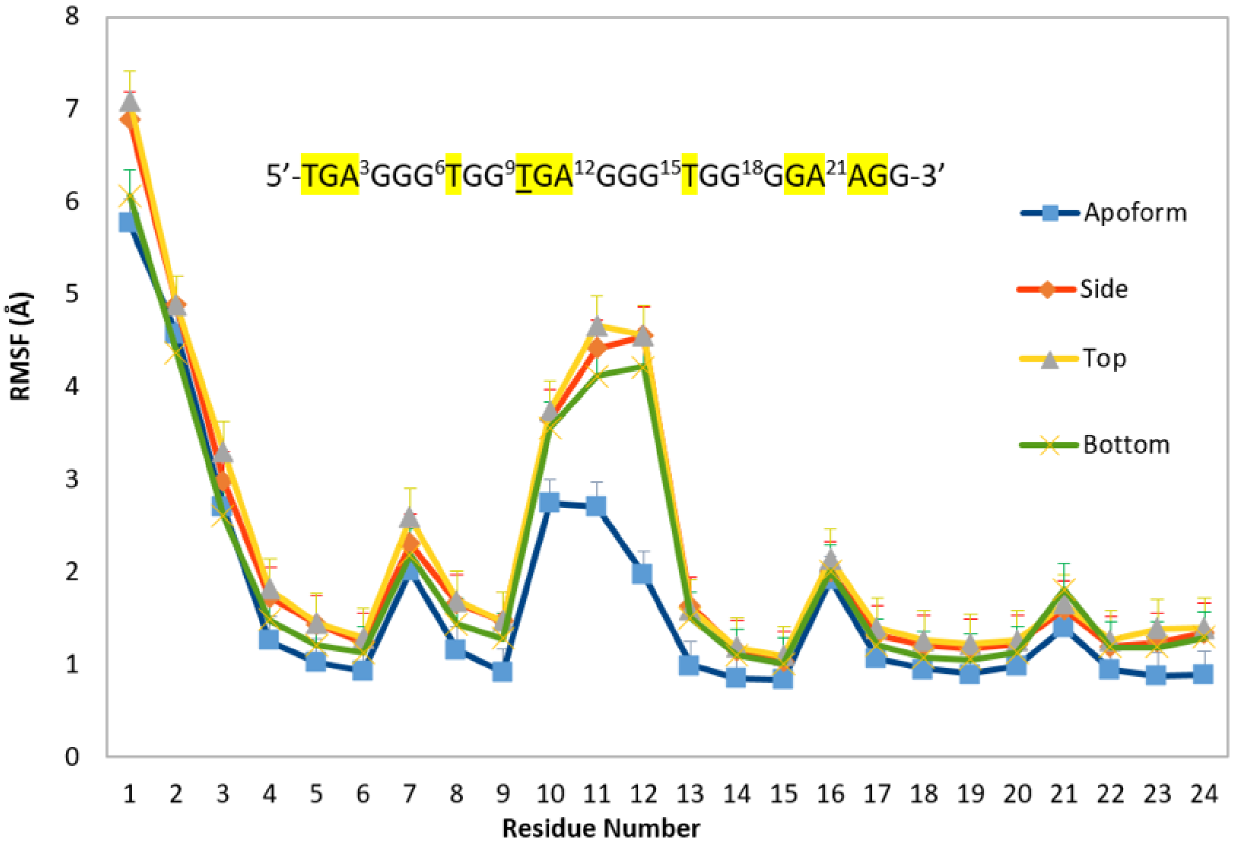


Figure 11. The Pu24 G-quadruplex RMSF by residue in the apo state and the ligand binding states (side, top, and bottom). The RMSF for the apo form is the average RMSF observed over 4 trajectories while the RMSF for the side, top, and bottom clusters are calculated from the individual complex clusters. Standard deviations are provided in matching colors. Non-G-tetrad residues are highlighted in yellow.

Discussion

DBD1, a Pu27 G-quadruplex stabilizer, is a novel compound that can selectively inhibit MYC transcription to induce apoptosis in cancer cells with minimal toxicity to normal human mononuclear cells. While the Pu27 structure has not yet been solved, the structure of the two derivatives with a different ligand (Pu22/DC-34 and Pu24/Phen-DC3), have been obtained. The structures of the two derivative G-quadruplexes are very similar and primarily differ in the 3' region; the extended 3'-end of Pu24 is connected to the bottom G-triad plane and restricts bottom-binding of Phen-DC3 to Pu24 while the truncated 3'-end of Pu22 allows for bottom-binding of DC-34 to Pu22. For the top binding modes of our MD structure (Pu24/DBD1) and the NMR solved Pu22/DC-34 complex, both of the respective ligands are positioned on top of their respective first G-tetrads and intercalation is observed. Not only does this support our findings that the binding of DBD1 to the top of the G-quadruplex is favorable, but it also validates our computational methods. Comparison of our bottom binding representative of the MD structure against the bottom binding 5W77 structure (**Figure 5**) showed that DC-34 is shown deeper within its bottom binding pocket, whereas DBD1 is more withdrawn. We attribute this binding difference to likely being due to the closed loop-like structure in the 3' region that is present in Pu24 but not present in Pu22. This difference has been already observed in the bottom-binding modes of Pu24/Phen-DC3 and Pu22/DC-34 complexes: while there is no bottom binding for Phen-DC3, the bottom binding mode of DC34 is very similar to its top binding mode. Of course, this difference in the two experimental complex structures could also be attributed to the difference in the ligand structures between Phen-DC3 and DC-34. However, our simulations use DBD1 in complex with Pu24 instead which negates much of the structural differences between the two ligands; the scaffolds of the two ligands are exactly the same with the only difference in the entire ligand structures being a methyl in DBD1 being replaced with a trifluoromethyl in DC-34. Both ligands were also shown to have similar K_d values, thus rooting out the possibility indicating that the ligand difference contributes to the observed bottom binding mode differences in the two experimental complex structures.

A recent synthetic biological study by Sengupta et al have shown, across multiple cell lines, that Pu24 induces significantly greater c-MYC promoter activity than Pu22, while inducing similar levels of promoter activity as the wild-type Pu27[30]. Their analyses of the effects of a kinase, NM23-H2, on Pu27 further showed that NM23-H2 promotes Pu27 forming the same conformation as Pu24, in order for G-quadruplex unfolding to initiate transcription[30]. We propose that the observed biological differences between Pu24 and Pu22 are likely due to this structural difference in the 3’ region, which leads to different unfolding efficiency by the G-quadruplex unfolding factor. While Pu22 has a flexible 3’ tail, Pu24 has a closed 3’ loop-like structure due to the G-triad formed by G20, A22 and G23. This speculation is supported by our observation that two highly similar ligands, DBD1 and DC34, had very different bottom binding modes when complexed with Pu24 and Pu22. Given that Pu24 more closely resembles the wild-type Pu27’s biological activity as compared to Pu22, the Pu24 G-quadruplex may be more insightful than the Pu22 G-quadruplex in predicting ligand effects in the c-MYC Pu27 G-quadruplex.

Our clustering and MMPBSA analyses suggest that the top binding mode is the most thermodynamically favorable amongst the three binding modes. This coincides with several other studies that show the top binding mode being the most thermodynamically favorable for certain G-quadruplex systems[83, 84]. The lesser energetic favorability exhibited by the side and bottom binding poses suggests that they might intermediate states. The side binding pose is not yet observed in either the 2MGN or 5W77 structures but has been observed in duplex structures[83]. Major contributors to the binding energy were van der Waals forces, which had -52.0±0.3 kcal/mol for the top binding mode, -14.7±1.0 kcal/mol for the bottom binding mode, and -14.3±0.5 kcal/mol for the side binding mode, and PB_{TOT}, which had -25.9±0.4 kcal/mol for the top binding mode, -4.5±0.8 kcal/mol for the bottom binding mode, and -4.5±0.5 kcal/mol for the side binding mode. Overall binding energies of the representative trajectories also showed that the top binding pose was the most stable, as the two trajectories that ended in the top binding pose exhibited MMPBSA values of -25 kcal/mol in comparison to the bottom or side binding MMPBSA values of -10 kcal/mol. Additionally, intercalation at the 5’-end of the G-quadruplex structure was observed for the top binding pose. Thus, we suggest that the ability for the ligand to bind in a planar orientation relative to the G-quadruplex is more energetically favorable, thus making the top binding mode the most favorable.

Further analysis of the kinetic relationship between the three states using MSM analysis revealed three parallel binding pathways to the top binding mode, the most thermodynamically stable state. Our

MSM analysis follows a similar procedure to that of our previous work[65] which also examined the binding pathways and kinetic information of G-quadruplex structures. Because we chose to cluster into a handful of “macrostates” directly during our MSM analysis and skipped over the experimentally unverifiable thousand “microstates”, the expected convergence time of the implied timescales should be significantly greater than that of a model with a greater number of clusters. This results in a coarser grained model that trades finer detail for greater experimental testability and easier human understanding[42]. It is likely that directly clustering into “macrostates” still maintains the integrity of the MSM as verification through the Chapman-Kolmogorov test (**Figure S11**) indicates that the model closely resembles the observed simulation data.

The interstate fluxes (**Figure S13**) indicate that the favored transition pathway is from unbound to top binding, though the pathways from unbound to side binding to top binding and unbound to bottom binding to side binding to top binding play a lesser but non-negligible role. This is supported by the mean-first passage time calculations which indicate that the unbound directly to top binding pathway is the fastest (1.6 ± 0.2 μs) while the other two pathways are several microseconds slower. The pathway from the unbound state directly to the final (top) binding state and the other two pathways that selectively bind to the side or the bottom pockets follows the conformational selection mechanism based on the comparison of the DNA conformation distribution before and after ligand binding (**Figure S19**). Huang and coworkers [44] have developed systematic approaches to estimate the abundance of induced-fit and conformational selection pathways in molecular recognition systems. Using their approach, they found out that the recognition mechanism of the choline binding protein to be ~90% conformation selection and ~10% of induced-fit under experimental conditions. In contrast, for this DNA-ligand system, there is no induced-fit and ~100% conformation selection. This difference between protein-ligand and DNA-ligand recognition appears be consistent with the note that protein is typically more plastic than DNA polymer.

We attempted to improve the DBD1 compound based on the insights from our MMPBSA analysis that the top binding mode was the most thermodynamically stable as well as the greatest contributors being from van der Waals and hydrophobic forces. Our hypothesis was that modifications to the pharmacophore, the planar 3-ring structure (**Figure 1C**), would improve the binding stability. Four sites S1-S4 were also substituted with additional atomic groups (S1 and S2 = H, F, Cl, OMe, NMe2 / S3 and S4 = H, F, Cl, NH2, CF3) which produced a library of 625 (5x5x5) new DBD1 derivatives (**Figure 12**).

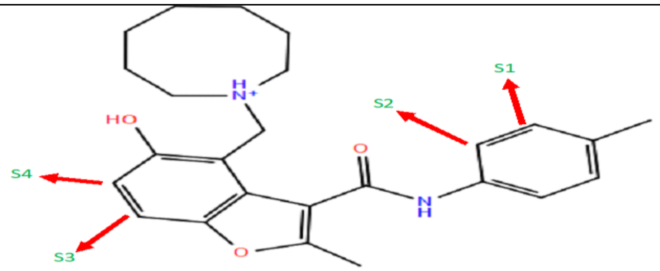
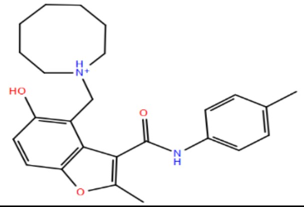
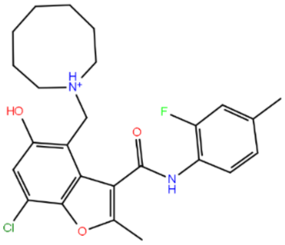
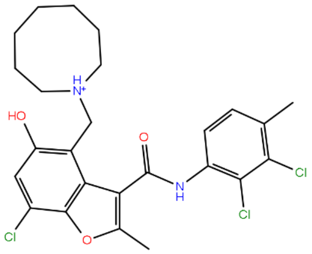
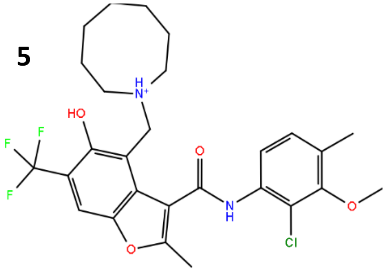
<table>
<tr><td colspan="3"> S1-S2 =H, F, Cl, OMe, NMe2
S3-S4=H, F, Cl, NH2, CF3</td></tr>
<tr><td><b>Chemical Structure of DBD1: Substitution Sites Marked</b></td><td colspan="2"><b>Docking Score (Kcal/mol)</b></td></tr>
<tr><td> </td><td colspan="2">-8.891</td></tr>
<tr><td><b>Chemical Structure of New Ligand</b></td><td><b>Docking Score</b></td><td><b>△DBD1</b></td></tr>
<tr><td><b>2</b>  </td><td>-9.255</td><td>-0.364</td></tr>
<tr><td><b>4</b>  </td><td>-9.487</td><td>-0.596</td></tr>
<tr><td><b>5</b>  </td><td>-9.565</td><td>-0.674</td></tr>
</table>

Figure 12. Chemical structure of DBD1 and 3 derivatives identified through virtual screening, including docking scores. For DBD1, red arrows indicate substitution sites for the derivatives displayed in this table. Substitution sites (S1-S2) contain either H, F, Cl, OMe and NMe2 atoms/groups and substitution sites (S3-S4) contain either H, F, Cl, NH2 and CF3 atoms/groups. For each derivative, docking scores are provided as well as the difference in docking scores compared with DBD1 reference ligand.

59 derivates from this library (**Figure S21)** showed slightly greater improvement in docking score compared to the reference ligand DBD1 to the G-quadruplex structure (the top-binding mode). The simplest derivative (Ligand 2) with one fluorine and one chlorine atom showed slight improvement in the Glide XP docking score [85](-0.364 kcal/mol improvement). A tri-substituted derivative (Ligand 4) containing three chlorine atoms showed a slightly more improved docking score (-0.596 kcal/mol improvement). Constitutively substituting sites S1-S4 with chlorine atoms (Ligand 30) shows slight improvement in the docking score (-0.524 kcal/mol improvement). Another tri-substituted ligand (Ligand 5) containing a CF_3 and OMe groups and a chlorine atom showed a more improved docking score (-0.674 kcal/mol improvement). The results imply that both halide and non-halide substitutions mildly improve the docking scores and binding affinities of DBD1 derivatives. Our observation is consistent with the original modification of DBD1 (D089-0563**)** to DC-34 by replacing a methyl with a trifluoromethyl, which slightly improved the binding affinity by about three fold from K_d=4.5 ± 1.4 μM to 1.4 ± 1.2 μM.[8,10]

Conclusion

DBD1 (D089-0563) is a novel G-quadruplex stabilizer that can selectively inhibit c-MYC, an oncogene, transcription to induce apoptosis in cancer cells by binding to the c-MYC G-quadruplex, Pu27. However, the structure of Pu27 is difficult to resolve due to its polymorphic and dynamic nature. Several derivatives with reduced morphologies have been created (Pu22, Pu24) to study their activity and structural differences. The complex of Pu24 with Phen-DC, structurally dissimilar to DBD1, as well as the complex of Pu22 with DC-34, an analog of DBD1 that replaces a methyl with a trifluoromethyl, have been recently solved, showing significant differences in the 3' region. However, the complex of Pu24 with DBD1 is not available, preventing the understanding of the working mechanism of DBD1 which is needed for further optimization. We performed free ligand simulations of DBD1 to Pu24 and subsequent clustering analysis, which identified three binding modes (top, side and bottom); interestingly, the top binding mode is fairly similar to the two experimental complexed structures, however the bottom binding mode of the Pu24 with DBD1 is different from DC34 complexed with Pu22. This suggests that the structural difference in the 3' region of the two G-quadruplexes leads to the difference in binding behavior of DBD1 at the 3' sites, which might also lead to the difference in the binding behavior of proteins that are involved in promoter activity to different G-quadruplexes. This difference in binding behavior may explain the difference in promoter activity. Through our MMPBSA calculations, the top binding pose was identified as the thermodynamically most favorable binding mode. Our MSM analysis showed multiple parallel pathways leading to the top binding mode, including a direct pathway and pathways involving intermediate states (side and bottom). These binding pathways suggest the conformational selection binding theory. In the top binding mode, the core part of the ligand stacks on top of the G-tetrad of the G-quadruplex and the stacking explains the van der Waals and hydrophobic interactions contributing the most to the stability of the top binding mode. Thus, we hypothesize that by increasing the core part by incorporating halogen substitutions might increase the binding interactions, thereby increasing its drug properties and stabilizing activity for cancer treatments. 625 D089-0563 derivatives were designed and docked; 59 of these showed slightly improved docking scores.

Supporting Information

Included in the supporting documents are the G-quadruplex sequences, the last snapshots, the most abundant receptor-ligand complexes from clustering analysis, superposition of the most abundant conformations from the MD simulations of the receptor with the ligands, histograms of the protein ligand interactions, a principal component analysis with RMSF, and the complete set of ligand torsion plots for all rotatable bonds throughout the MD trajectory, AMBER GAFF2 force field of the ligand DBD1 (+1) in Mol2 format.

Acknowledgements

This work was supported by Rowan CSM SEED grant and the National Science Foundation under Grants NSF RUI-1904797/ACI-1429467 and XSEDE MCB 170088.

References

1. K. Suntharalingam, A. J. P. White and R. Vilar, *Inorganic Chemistry*, 2009, 48, 9427-9435.
2. S.-T. D. Hsu, P. Varnai, A. Bugaut, A. P. Reszka, S. Neidle and S. Balasubramanian, *Journal of the American Chemical Society*, 2009, 131, 13399-13409.
3. T. Agarwal, M. K. Lalwani, S. Kumar, S. Roy, T. K. Chakraborty, S. Sivasubbu and S. Maiti, *Biochemistry*, 2011, 53, 1117-1124.
4. S. Harikrishna, S. Kotaru and P. I. Pradeepkumar, *Molecular Biosystems*, 2017, 13, 1458-1468.
5. P. Agrawal, E. Hatzakis, K. Guo, M. Carver and D. Yang, *Nucleic Acids Res*, 2013, 41, 10584-10592.
6. A. T. Phan, Y. S. Modi and D. J. Patel, *Journal of the American Chemical Society*, 2004, 126, 8710-8716.
7. E. Ruggiero and S. N. Richter, *Nucleic Acids Res*, 2018, 46, 3270-3283.
8. A. Ambrus, D. Chen, J. Dai, R. A. Jones and D. Yang, *Biochemistry*, 2005, 44, 2048-2058.
9. A. Tawani, S. K. Mishra and A. Kumar, *Scientific Reports*, 2017, 7.
10. S. Neidle, *Journal of Medicinal Chemistry*, 2016, 59, 5987-6011.
11. P. H. Watson, J. R. Safneck, K. Le, D. Dubik and R. P. C. Shiu, *JNCI: Journal of the National Cancer Institute*, 1993, 85, 902-907.
12. D. Hawksworth, L. Ravindranath, Y. Chen, B. Furusato, I. A. Sesterhenn, D. G. McLeod, S. Srivastava and G. Petrovics, *Prostate Cancer And Prostatic Diseases*, 2010, 13, 311.
13. U. R. Rapp, C. Korn, F. Ceteci, C. Karreman, K. Luetkenhaus, V. Serafin, E. Zanucco, I. Castro and T. Potapenko, *PLoS One*, 2009, 4, e6029.
14. H. J Wu, *Nihon Sanka Fujinka Gakkai zasshi*, 1996, 48, 515-521.
15. D. R. Smith, T. Myint and H. S. Goh, *Br J Cancer*, 1993, 68, 407-413.
16. L. F. Barr, S. E. Campbell, G. B. Diette, E. W. Gabrielson, S. Kim, H. Shim and C. V. Dang, *Cancer Research*, 2000, 60, 143.
17. I. Magrath, in *Advances in Cancer Research*, eds. G. F. Vande Woude and G. Klein, Academic Press, 1990, vol. 55, pp. 133-270.
18. B. G. Kim, H. M. Evans, D. N. Dubins and T. V. Chalikian, *Biochemistry*, 2011, 54, 3420-3430.
19. P. V. L. Boddupally, S. Hahn, C. Beman, B. De, T. A. Brooks, V. Gokhale and L. H. Hurley, *Journal of Medicinal Chemistry*, 2012, 55, 6076-6086.
20. R. I. Mathad, E. Hatzakis, J. Dai and D. Yang, *Nucleic Acids Res*, 2011, 39, 9023-9033.
21. M. Cooney, G. Czernuszewicz, E. H. Postel, S. J. Flint and M. E. Hogan, *Science*, 1988, 241, 456.

22. T. L. Davis, A. B. Firulli and A. J. Kinniburgh, *Proc Natl Acad Sci U S A*, 1989, 86, 9682-9686.
23. J. T. Davis, *Angewandte Chemie International Edition*, 2004, 43, 668-698.
24. W. J. Chung, B. Heddi, F. Hamon, M. P. Teulade-Fichou and A. T. Phan, *Angewandte Chemie-International Edition*, 2014, 53, 999-1002.
25. N. Deng, L. Wickstrom, P. Cieplak, C. Lin and D. Yang, *J Phys Chem B*, 2017, 121, 10484-10497.
26. T. Che, Y. Q. Wang, Z. L. Huang, J. H. Tan, Z. S. Huang and S. B. Chen, *Molecules*, 2018, 23.
27. K. M. Felsenstein, L. B. Saunders, J. K. Simmons, E. Leon, D. R. Calabrese, S. Zhang, A. Michalowski, P. Gareiss, B. A. Mock and J. S. Schneekloth, *ACS Chemical Biology*, 2011, 11, 139-148.
28. S. P. Pany, P. Bommisetti, K. V. Diveshkumar and P. I. Pradeepkumar, *Org Biomol Chem*, 2016, 14, 5779-5793.
29. D. R. Calabrese, X. Chen, E. C. Leon, S. M. Gaikwad, Z. Phyo, W. M. Hewitt, S. Alden, T. A. Hilimire, F. He, A. M. Michalowski, J. K. Simmons, L. B. Saunders, S. Zhang, D. Connors, K. J. Walters, B. A. Mock and J. S. Schneekloth, Jr., *Nature communications*, 2018, 9, 4229.
30. P. Sengupta, A. Bhattacharya, G. Sa, T. Das and S. Chatterjee, *Biochemistry*, 2019, 58, 1975-1991.
31. H. Zhu, S. Xiao and H. Liang, *PLoS One*, 2013, 8, e71380.
32. J. Sponer and N. a. Spacková, *Methods*, 2007, 43, 278-290.
33. J. Sponer, X. Cang and T. E. Cheatham, 3rd, *Methods*, 2012, 57, 25-39.
34. H.-J. Kang and H.-J. Park, *Biochemistry*, 2009, 48, 7392-7398.
35. D. L. Ma, D. S. Chan, W. C. Fu, H. Z. He, H. Yang, S. C. Yan and C. H. Leung, *PLoS One*, 2012, 7, e43278.
36. J. Bhat, S. Mondal, P. Sengupta and S. Chatterjee, *ACS Omega*, 2017, 2, 4382-4397.
37. J. Dai, M. Carver, L. H. Hurley and D. Yang, *Journal of the American Chemical Society*, 2011, 133, 17673-17680.
38. O. Buket, L. Clement and Y. DanZhou, *Sci China Chem*, 2014, 57, 1605-1614.
39. D. E. Koshland Jr, *Angewandte Chemie International Edition in English*, 1995, 33, 2375-2378.
40. J.-P. Changeux and S. Edelstein, *F1000 Biol Rep*, 2011, 3, 19-19.
41. H.-J. Kang and H.-J. Park, *J Comput Aided Mol Des*, 2014, 29, 339-348.
42. V. S. Pande, K. Beauchamp and G. R. Bowman, *Methods*, 2010, 52, 99-105.
43. D. A. Silva, G. R. Bowman, A. Sosa-Peinado and X. H. Huang, *Plos Computational Biology*, 2011, 7.
44. S. Gu, D. A. Silva, L. M. Meng, A. Yue and X. H. Huang, *Plos Computational Biology*, 2014, 10.
45. A. Siddiqui-Jain, C. L. Grand, D. J. Bearss and L. H. Hurley, *Proc Natl Acad Sci U S A*, 2002, 99, 11593-11598.
46. S. Burge, G. N. Parkinson, P. Hazel, A. K. Todd and S. Neidle, *Nucleic Acids Res*, 2006, 34, 5402-5415.
47. A. K. Todd, M. Johnston and S. Neidle, *Nucleic Acids Res*, 2005, 33, 2901-2907.
48. J. Seenisamy, E. M. Rezler, T. J. Powell, D. Tye, V. Gokhale, C. S. Joshi, A. Siddiqui-Jain and L. H. Hurley, *Journal of the American Chemical Society*, 2004, 126, 8702-8709.
49. A. T. Phan, V. Kuryavyi, H. Y. Gaw and D. J. Patel, *Nat Chem Biol*, 2005, 1, 167-173.
50. A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham, 3rd, C. A. Laughton and M. Orozco, *Biophysical journal*, 2007, 92, 3817-3829.
51. M. Krepl, M. Zgarbova, P. Stadlbauer, M. Otyepka, P. Banas, J. Koca, T. E. Cheatham, 3rd, P. Jurecka and J. Sponer, *Journal of chemical theory and computation*, 2012, 8, 2506-2520.
52. M. Zgarbova, F. J. Luque, J. Sponer, T. E. Cheatham, 3rd, M. Otyepka and P. Jurecka, *Journal of chemical theory and computation*, 2013, 9, 2339-2354.
53. M. Zgarbova, J. Sponer, M. Otyepka, T. E. Cheatham, 3rd, R. Galindo-Murillo and P. Jurecka, *Journal of chemical theory and computation*, 2015, 11, 5723-5736.

54. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, 79, 926-935.
55. I. S. Joung and T. E. Cheatham, 3rd, *The journal of physical chemistry. B*, 2008, 112, 9020-9041.
56. D. Case, R. M. Betz, D. S. Cerutti, T. Cheatham, T. Darden, R. Duke, T. J. Giese, H. Gohlke, A. Götz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.-S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko and P. A. Kollman, *Amber 16, University of California, San Francisco*, 2016.
57. C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *The Journal of Physical Chemistry*, 1993, 97, 10269-10280.
58. K. Mulholland and C. Wu, *Journal of chemical information and modeling*, 2016, 56, 2093-2102.
59. Z. H. Shen, K. A. Mulholland, Y. J. Zheng and C. Wu, *Journal of Molecular Modeling*, 2017, 23.
60. K. Mulholland, F. Siddiquei and C. Wu, *Physical Chemistry Chemical Physics*, 2017, 19, 18685-18694.
61. J. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comp. Phys.*, 1977, 23, 327-341.
62. U. Essmann, L. Perera, M. L. Berkowitz, T. A. Darden, H. Lee and L. G. Pedersen, *J. Comp. Phys.*, 1995, 103, 8577-8593.
63. P. Procacci and B. J. Berne, *Molec. Phys.*, 1994, 83, 255-272.
64. H.-J. Sullivan, C. Readmond, C. Radicella, V. Persad, T. J. Fasano and C. Wu, *ACS Omega*, 2018, 3, 14788-14806.
65. K. Mulholland, H.-J. Sullivan, J. Garner, J. Cai, B. Chen and C. Wu, *ACS chemical neuroscience.*, 2020, 11, 57-75.
66. D. Bhattacharyya, G. Mirihana Arachchilage and S. Basu, *Frontiers in Chemistry*, 2016, 4, 38.
67. W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics & Modelling*, 1996, 14, 33-38.
68. Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, C. Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, L.-P. Wang, Thomas J. Lane and Vijay S. Pande, *Biophysical journal*, 2015, 109, 1528-1532.
69. F. Pedregosa, Ga, #235, l. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, #201 and d. Duchesnay, *J. Mach. Learn. Res.*, 2011, 12, 2825-2830.
70. P. J. Rousseeuw, *Journal of Computational and Applied Mathematics*, 1987, 20, 53-65.
71. X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren and A. E. Mark, *Angew. Chem., Int. Ed.*, 1999, 38, 236-240.
72. J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte and F. Noé, *The Journal of Chemical Physics*, 2011, 134, 174105.
73. F. Noé, I. Horenko, C. Schütte and J. C. Smith, *The Journal of Chemical Physics*, 2007, 126, 155102.
74. G. Csardi and T. Nepusz, *InterJournal, Complex Systems*, 2006, 1695, 1--9.
75. P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. I. Cheatham, *Acc. Chem. Res*, 2000, 33, 889-897.
76. D. S. C. D.A. Case, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman, University of California, San Francisco, 2016.
77. J. Kongsted, P. Soderhjelm and U. Ryde, *J. Comput. Aided Mol. Des.*, 2009, 23, 395-409.

78. P. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. Case and T. Cheatham, *Acc. Chem. Res.*, 2000, 33, 889-897.
79. T. Hou, J. Wang, Y. Li and W. Wang, *J. Comput. Chem.*, 2010, 32, 866-877.
80. T. Hou, J. Wang, Y. Li and W. Wang, *J. Chem. Inf. Model.*, 2011, 51, 69-82.
81. L. Xu, H. Sun, Y. Li, J. Wang and T. Hou, *J. Phys. Chem. B.*, 2013, 117, 8408-8421.
82. H. Sun, Y. Li, S. Tian, L. Xub and T. Hou, *PCCP*, 2014, 16, 16719-16729.
83. B. Machireddy, H.-J. Sullivan and C. Wu, *Molecules (Basel, Switzerland)*, 2019, 24, 1010.
84. Z. Shen, K. A. Mulholland, Y. Zheng and C. Wu, *Journal of Molecular Modeling*, 2017, 23, 256.
85. D. Kitchen, H. Decornez, J. Furr and J. Bajorath, *Nat Rev Drug Discov*, 2004, 3, 935-949.