MANUSCRIPT TITLE

Molecular Dynamics Reveal Base Flipping as a Key Mechanism in Tc3a Transposase DNA Recognition

AUTHORS

Stephan L. Watkins [1,#a,*], and Helder I. Nakaya[1,2,3,*]

[1] Hospital Israelita Albert Einstein, São Paulo, SP, Brazil

[2] Department of Clinical and Toxicological Analyses, University of São Paulo, São Paulo, SP, Brazil

[3] Institut Pasteur de São Paulo, São Paulo, SP, Brazil

[*] Corresponding Authors
Email. helder.nakaya@einstein.br ,
Email. stephenlloydriggs@gmail.com, stephan.lloyd@ipt.br

[#a] Present Address. CSBL, University of São Paulo, INOVA Building, São Paulo, Brazil

Abstract

Transposons, or "jumping genes," are mobile genetic elements that reshape genomes. The Tc1/Mariner family, including Tc3a transposase, has been studied for decades, but key aspects of its DNA recognition and cleavage remain unresolved. Existing models propose a homo-dimeric process for DNA excision and integration, yet conflicting hypotheses exist about the initial steps of DNA cleavage and target site recognition. Here, we reveal a previously unrecognized base-flipping mechanism in Tc3a transposase that challenges long-standing models of transposon activity. Using molecular dynamics simulations, we demonstrate that N-terminal acetylation induces a structural shift in the transposase-DNA complex, forcing a nucleotide in the inverted repeat to flip outward. This base-flipping event alters the local DNA conformation, creating torsional strain that may facilitate transposon cleavage or integration. Unlike prior models, which assumed static DNA interactions, our findings indicate that transposase binding itself actively reshapes DNA structure. We further show that this mechanism is dependent on the bipartite helix-turn-helix domains, which differentially contribute to DNA stabilization and recognition. Our simulations reveal that the processed N-terminus plays an unexpected role in modulating DNA binding affinity, contradicting previous assumptions that it was structurally inconsequential. Together, the two helix-turn-helix motifs act to propagate the force caused by base flipping along one direction of the cognate DNA double helix axis. The fact that previous structural studies lacked this amino acid modification may explain why this mechanism was overlooked. These results provide a new framework for understanding transposase-DNA interactions, highlighting how a single molecular modification can trigger major DNA rearrangements. This discovery not only redefines the Tc3a transposition process but also calls for a reassessment of long-standing mariner family transposon models.

Keywords

Transposase, HTH-helix-turn-helix, Molecular dynamics, Bipartite

Introduction

Transposons are small stretches of DNA capable of "jumping" to different regions of an organism's genome, profoundly influencing genetic diversity and evolution [1, 2]. These mobile elements are now classified into various families, each characterized by inverted terminal repeats (ITRs) flanking a central region (1200–4000 bp) that encodes a transposase, the enzyme responsible for their mobility [3, 4]. Transposons have been identified in every organism studied, ranging from inactive pseudo-transposons to fully functional elements [5]. Their ability to facilitate horizontal gene transfer across kingdoms—such as between Protozoa and Animalia, or even between Oomycetes and Bryophytes—underscores their evolutionary significance [6]. Moreover, their simplicity has made them invaluable tools in genetics, with applications in model organisms like Caenorhabditis elegans, Drosophila melanogaster, and Danio rerio, as well as potential candidates for gene therapy [7–11].

This study focuses on the Tc1/Mariner superfamily of transposons, first identified in C. elegans in 1989 [6, 12]. The Mariner superfamily encompasses over 1000 members, predominantly found in worms and insects, although some have been integrated into the genomes of humans and other vertebrates [13, 14]. A hallmark of this superfamily is its simplicity, a single transposase mediates the entire transposition process, guided by the ITRs that flank each transposon. Structurally, these proteins share a conserved N-terminal helix-turn-helix (HTH) DNA recognition domain, a short linker chain followed by a second HTH and a catalytic C-terminal core. This catalytic core is conserved throughout Tc1/mariners, consisting of a DDE/D Mg^{2+} binding site responsible for the entire excision and insertion processes defining the transposon jumping [15, 16]. Transposition is thought to occur via a homo-dimeric mechanism in which the transposase excises the transposon at the ITR ends and inserts it at TATA-like sequences elsewhere in the genome. However, key questions about the initial DNA cleavage and target recognition remain unresolved [17, 18]. These include several differing models of how the initial DNA cleavage occurs, ranging from complex multimerization of 4 transposase into a large synaptic complex

98 to single tranposase cleavage at ends, then dimerization after cleavage. Intriguingly, the structural
99 homology of Tc1/Mariner transposases to transcription factors in higher organisms and restriction
100 enzymes in prokaryotes suggests a shared evolutionary history, likely influenced by horizontal gene
101 transfer [7, 19, 20].

102 Here, we investigated the ITR recognition of the Tc3a transposase to provide new insights into DNA
103 binding and specificity in this protein family. Despite high structural similarity, Tc1/Mariner proteins
104 showed divergent sequence homology, making them an excellent model for studying the balance between
105 specificity and general DNA recognition [20–22]. Our aim was to uncover the kinetics and mechanisms
106 underlying these interactions, shedding light on DNA recognition processes relevant to transcription
107 factors and DNA-binding proteins in diverse organisms, including Homo sapiens. Furthermore, we
108 explored a novel DNA-binding modality that appears to be a foundational mechanism in these
109 transposons, with implications for understanding DNA and RNA interactions across vertebrates, plants,
110 and mammals. This research also addresses a critical gap in engineering synthetic DNA-recognition
111 proteins for biotechnological and therapeutic applications.

112

113 Materials and Methods

114

115 Starting Structural Model.

116 Our initial structure was retrieved from the Protein Data Bank PDB id=1U78, and modified accordingly
117 [23]. Based on normal C. elegans and H. sapiens protein processing, the N terminal methionine was
118 removed [24], and terminal Proline was modified to include a terminal acetylation. From the PDB DNA,
119 the bound DNA was modified to include the Tc3 ITR sequence, 5'-
120 TGTGGGAAAGTTCTATAGGACCCCCCCG-3' and complement, 3'-

121 CACACCCTTTCAAGATATCCTGGGGGGG-5', which encompasses the entire N-terminal protein
122 recognition portion, using Pymol. Gromacs-GPU 5.1.15 was initially used, with an all-atom Amber
123 ff99SB-ILDN force field. Our Resulting model was centered in a rectangular box with dimensions
124 100x56x180 Å, solvated with 31726 TIP3P water molecules, 1 Mg2+ ion (0.002 M), 91 Na+ ions (0.140
125 M), and 30 K+ ions (0.040 M), with counter Cl- ions for each positive ion to mimic a minimal cellular
126 salt concentration. The resultant box gave a 13 Å solvent layer to the Z axis, 40 Å to the X and 60 Å in
127 the Y dimension [25].
128
129

130 Molecular Dynamics Model Equilibration

131 The resulting model was energy minimized using steepest descent, and equilibrated to 310 Celsius, 1
132 atmosphere by first generating velocities at 290 Celsius for 2 ns, then continuing runs for 5 ns at 310
133 Celsius, with Nose-Hoover temperature coupling, and Parrinello-Rahman Pressure coupling. A resulting
134 structural model was then further equilibrated using unrestrained MD with three temperature, pressure
135 coupling groups DNA, Protein and water plus ions, for 20 nanoseconds. For further experiments, this
136 resulting simulation was allowed to run for an additional 50 nanoseconds to generate adequate sampling
137 space for starting pulled simulations. All pulled simulations were done using Gromacs GPU version
138 2023.2 on 4 H100 GPUs and 48 core Intel 2.3 GHz servers on the Brazilian National Computing
139 Laboratory (LNCC), or 16 cores, 1 H100 GPU on CENAPAD Brazilian super computers, using all-atom
140 Amber ff99SB-ILDN force field, long-range cutoffs set to 1.4 nanometers, neighbor list update set to 2.4
141 picoseconds, and Particle Mesh Ewald long range electrostatics, v-rescale temperature coupling, and c-
142 rescale Pressure coupling for GPU use [26]. All models were then run an additional 50 ns to determine
143 if any further fluctuations or changes occur, and for bound DNA interaction analysis. This process was

144 repeated for models having an N-terminal modification with cleaved methionine and no acetyl group,
145 the terminal three residues remove, and the methionine followed by alanine and glycine, and no
146 processing. Hydrogen bonding or structural comparisons were analyzed using Gromacs, APBS software
147 and in the software VMD [27, 28]. Model comparisons were done using Gromacs, VMD and Pymol
148 [29, 30].

149 Free Energy Calculations

150 Gibbs free energy was conducted according to published PMF protocols using the above models [31,
151 32]. Briefly, a 50 nanosecond segment was used to generate starting models with time points between 5
152 and 50 nanoseconds chosen using a random time point generator. This was done 4 times, for the 4
153 separate N-terminal models systems. First 16 time points were used to create a starting point with the
154 first HTH comprising the first 46 amino acids, HTH1, as the applied pulled force, with the DNA
155 restrained at the 3'-G53 and 3'-C28 terminal Oxygen atoms with 2000 kJ/mol. Using Umbrella, and
156 distance, a pull force was applied of 2000 kJ/mol nm^2, pull rate of 1 nm per nanosecond to the HTH1
157 center of mass (COM), with a pull tolerance of 10^{-06} kJ/mol/ps, and the simulation was run with with V-
158 rescale temperature coupling, C-rescale Pressure coupling, long-range cutoffs set to 1.4 nanometers,
159 neighbor list update set to 2.4 picoseconds, Particle Mesh Ewald long range electrostatics, using an all-
160 atom Amber ff99SB-ILDN force field with the HTH1 over 7.5-8 nanoseconds. Bound hydrogen bonds
161 were restrained using the SHAKE algorithm to constrain bond vibrations so we could use a time step of
162 2 fs [33]. This was repeated using amino acids 47-57 for the connecting loop domain as COM applied
163 force, and again for the second HTH comprised of amino acids 58-104 , HTH2. This resulted in 16
164 pulled simulations for each domain. A final set of 26 pulled simulations was then conducted, again using
165 randomized time point selection, with HTH1, the loop domain, and HTH2 all pulled from their COM
166 simultaneously. This was repeated for each of the described models, using the same three pull groups,

167 and in all frames saved at 1 ps intervals. Data was analyzed in Gromacs using the Wham method with
168 only the the three simultaneously pulled runs used for energy plots, and all simulations used for structural
169 comparisons [34, 35].

170 Further Analysis

171 Further analysis was done using Gromacs analysis tools. For Root mean square fluctuations with the
172 bipartite, 300 ps in the terminal portion of the pulled simulations was compared to 300 ps in the bound
173 state and used for figures. For the RMSF of the ITR DNA, the entire pull simulations were used. In
174 both, the presented data is the mean average of all simulations. For dynamic hydrogen bonding and
175 density plots the mean average was calculated from all data across the entire equilibration runs, truncating
176 the first 5 ns, excluding the equilibration temperature and pressure normalization [36]. For transition
177 hydrogen bonding and distance, averages were produced from the entire pulled simulation sets. A
178 representative distance plot was used in Fig 8 as averaged plots vary significantly in starting time points
179 or duration, causing the data to be lost in noise from averaged plotting of means. Model snapshots, and
180 trajectories were further analyzed, and graphic representations or figures made using VMD, Pymol,
181 Qtiplot, Gimp and Weblogo [37, 38].

182 Protein Database Analysis

183 Protein Database Analysis was conducted as follows. The entire Uniprot protein database was
184 downloaded as fasta files, and separated into Swiss-protein, and Tremble data sets. These were parsed
185 with python scripts to tabulate and bin the fasta files containing N-terminal Pro-Arg. Further analysis
186 was done using the resulting fasta files to assign percentages, classified according to organisms as listed
187 in the Uniprot database, and generate tabulated data using Pfam, Interpro and GO based protein
188 functional identifiers to assign protein function [39–41]. Statistics were done based on direct DNA or
189 RNA binding protein identifiers. Parsed H.sapiens data using the N-terminal Pro-Arg, Pro-His, Pro-Gln,

190 Ala-Leu, Ser-Arg, and Arg-Gln to extract all fastas, was reduced based on annotation and to remove
191 redundant transcript variants, and enrichment analysis preformed with DAVID and MOET [42, 43].
192
193

194 Results

195 Initial Differences from X-ray structures

196 Initial equilibrated model structure hydrogen bonding varies from the X-ray structure, mostly through
197 expected thermal expansion from the cryogenic determined atomic positions. With x-ray or EM models
198 data is usually collected at -180 Celsius, (Fig 1). Main differences were Tyr46 made no hydrogen bonds
199 with the DNA in any simulations, along with the Ser6 and Ala6 backbone in the x-ray structure. Our
200 initial model also showed slightly different hydrogen bonding between the two helix-turn-helix (HTH)
201 domains, referred to as HTH1 for the N-terminal and HTH2 for the C-terminus, as well as the connecting
202 loop analyzed over 50 ns, from a 100 ns simulation. This loop sits more deeply in the widened minor
203 groove, and makes different hydrogen bonds due to dynamics not completely included in the static
204 structure [23, 44]. Residues are number starting with the cleaved N-terminus methionine or acetyl group
205 as 1. Beginning with the acetylated-Pro, the residue is almost stationary in position with respect to the
206 ITR DNA. The Acetyl oxygen forms a optimized repulsive interaction at 3.4 Å with the ring N3 of A12,
207 forming a perfect 109° angle with the sugar O4' of T13, and O4' to N1 of A43 angle at 121.6º and 4.0
208 Å. A slight 3.3 Å hydrogen bonding to the almost neutral H2 of A43 may help stabilize this reaction,
209 although the bonding only occurs transiently 6% of occupied time, or with a lifetime of only 2.4 ps, along
210 with 3% and 3% respectively for H4' and H5'1 of T13 sugar ring, each with a 1-1.2 ps lifetime or less.
211 This position is more reinforced by the proline's almost complete hydrophobicity due to the neutralizing
212 nitrogen modification, pushing the residue into the center of the DNA, and being completely solvent

213 exposed on one side. Residue Arg3 forms an antiparallel base pair through NH2 hydrogen and the sugar
214 of T11 O4', trapping the base O2 which is also bound by the hydrogen of the NE. This places the Arg
215 main-chain N within 2.5 Å of the complementary base, A45, while base N3, formed a bond occupied
216 68.3% of the time, and a salt bridge with O1 of T13 phosphate, mediated by a fixed H2O, 50% occupied
217 through simulations.
218
219 **Fig 1. Initial Steered Simulation Model. A]** A representation of the unit cell, with the bipartite centered
220 in the middle, ions K, purple, Na, blue, water is not shown. The force was applied along the long axis,
221 and simulations run until half the protein entered the periodic wall. **B]** Cartoon representation of the
222 equilibrated starting model, N-terminus and HTH1 is to the right. **C]** The sequence of the dsDNA, aligned
223 to the model in B], larger letters indicate a contact with the bipartite protein.
224
225 Together this forces a contraction of the minor groove on one side towards HTH1, and a widening,
226 aided by Arg54 immediately below the modified Pro2 in the adjacent direction. These bonds, Arg54
227 NH2 hydrogen with A43 base N3, Arg54 NE hydrogen with A43 ribose O4', and Arg54 NE hydrogen
228 with T42 base O2 occupied extensively in simulations. This interaction pulls slightly and stabilizes A43,
229 opposite T13. Together, the acetylated N-terminal Pro-Arg3 and Arg54 cause the T13 to adopt a flipped
230 out conformation rotating 120-140 degrees into the solvent. This conformation, not observed in x-ray
231 data, is maintained through the duration of simulations. A net effect is to cause extensive torsion on one
232 strand of the DNA in the duplex, that would translate to coiling in longer extensions of DNA fragments
233 [16, 45, 46]. The charge distributions, and structure of the region are shown in Fig 2, highlighting the
234 effective use of hydrophobic force and stabilizing hydrogen bonding that can allow base flipping in small
235 HTH containing DNA binding proteins. All hydrogen bonds lasting longer than 10 ps, and having an
236 occupancy higher than 40% for longer bound simulations are presented in Table 1.
237

Fig 2. N-terminus Binding Site. A] Cartoon of the binding site for the N-terminus, including Arg54. Bases are labeled with 1 letter DNA code and position number (Fig.1). Hydrogen bonds are colored as yellow dotted lines. **B]** Same orientation as A], DNA shown as space filling model with APBS calculated partial charges shown according to scale as red, negative and blue, positive. Protein is shown as stick model, colored by atom type, red, oxygen, blue nitrogen, carbon light blue, hydrogen white. A hole formed through the DNA from flipped T13 is occupied by the Acetyl group.

Table 1 Long Lived High Occupancy Protein-DNA Hydrogen Bonds

Functional Group		
H-Donor	H-Acceptor	Percent Occupancy
ARG3-NE	T11-O2	99.00%
ARG3-NH2	T11-O2	78.34%
ARG3-NH2	T11-O4’	7.85%
ARG3-NH1	A12-O4’	21.06%
SER24-OG	G5-O1P	94.95%
SER24-OG	G6-O1P	11.65%
HIS26-NE2	G6-N7	80.85%
HIS26-NE2	G5-N7	6.07%
ARG30-NH1	G5-O1P	99.03%
ARG30-NH2	G5-O2P	98.53%
ARG34-NH1	G46-O1P	96.85%
SER35-OG	G46-O2P	99.10%
ARG36-NH1	G7-N7	99.70%
ARG36-NH2	G7-O6	93.87%
C9-N4	HIS37-ND1	52.39%
HIS37-NE2	G46-N7	22.52
HIS37-NE2	A45-N7	6.92%
CYS38-HG	A45-O5’	10.62%
CYS38-HG	G46-O2P	7.87%
ARG40-NH2	G7-O2P	81.53%
ARG40-NH1	T8-O2P	17.39%
SER52-OG	A44-O2P	44.89%
SER52-OG	A43-O3’	12.85%
LYS53-NZ	A16-O1P	5.45%
LYS53-NZ	A17-O1P	35.32%
ARG54-NH2	A43-N3	99.10%
ARG54-NH1	A43-O4’	11.29%
ARG54-NH1	A14-N3	7.6%

ARG54-NE	T42-O2	48.91%
ARG57-NE	T39-O2	32.69%
ARG57-NH1	C18-O2	26.77%
ARG57-NH2	T40-O1P	8.57%%
ARG58-NH2	T42-O2P	45.24%
ARG58-NH1	T41-O2P	35.44%
ARG58-NH2	C41-O3’	47.31%
LYS59-NZ	C18-O1P	21.67%
THR79-OG1	T31-O2P	76.06%
ARG81-NH1	G30-O2P	52.01%
ARG81-NH2	T31-O2P	48.69%
SER92-OG	T20-02P	47.24%
SER92-OG	T20-O5’	47.22%
LYS93-NZ	G32-O6	60.98%
LYS93-NZ	G33-O6	39.11%
ARG94-NH2	T21-O4	9.42%
ARG94-NH2	G34-N7	54.29%
THR95-OG1	T20-O2P	97.15%
LYS101-NZ	G33-O2P	36.19%
ARG102-NH1	G33-O2P	98.31%
ARG102-NH2	G33-O2P	98.27%
Backbone		
ARG3-N	A45-N3	68.01%
ARG3-N	T44-04’	9.80%
GLY4-N	A45-O4’	14.15%
LEU25-N	G6-O1P	76.63%
LEU25-N	G6-O2P	4.47%
HIS26-N	G6-02P	59.66%
SER35-N	G46-O2P	89.20%
GLY50-N	A45-O2P	27.22%
GLY50-N	A44-O1P	6.15%
THR51-N	A44-O1P	33.82%
SER52-N	A44-O1P	43.96%
ALA55-N	A43-O1P	11.94%
C41-H1’	PRO56-O	7,47%
ARG57-N	C41-O4’	21.04%
ARG58-N	C41-O2P	47.31%
LYS59-N	C18-O1P	7.97%
ALA60-N	T19-O1P	92.98%
ALA80-N	G32-O2P	72.31%
ALA91-N	T20-O1P	16.05%
SER92-N	T20-O1P	48.42%
SER92-N	T20-O2P	48.41%

248 List of all hydrogen bonds between Protein and DNA with higher than 10 ps, 10% occupancy across
249 extended simulation. Amino acids that meet the criteria once also have shorter bond occupancies listed.
250 Functional Group, atomic interactions from amino sidechain atoms, Backbone, atomic interactions from
251 amino mainchain. In both column 1, donor, column 2 acceptor, column 3 percent occupancy across 100
252 ns pre-equilibrated simulation.
253

254 Distortions in the DNA only slightly mirror x-ray structures of Tc3a, Mos1 or related Mariners. The
255 flipped T13 causes the phosphate of A14 to bend inwards towards the minor groove, however this torsion
256 causes the minor groove between G15 to C17 and C41 to A45 to adopt a wider opening. Respective
257 averages by base from G15 are 15.54 Å, with a maximum 19.34 Å, minimum 11.89 Å , to A17 with an
258 average of 13.26 Å , and maximum 15.71 Å, minimum 11.32 Å. The highest distortion occurs from
259 G15 and A16 paired with T44 and A45 on the opposite strand. Conversely, the opposite direction of the
260 flipped T13 causes the minor groove between DNA residues from C9 to A12 and C51 to G48 to contract.
261 Here the highest variation is on C10 and T11, with average, minimum and maximum of 10.26 Å , 6 Å,
262 15.20 Å and 12.33 Å, 6.56 Å, 15.51 Å respectively. Residues C9 and A12 have only minimums of 8.9
263 Å, with averages in normal B-DNA ranges. With both distortions, the widening in one direction and
264 contraction the opposite stretch, the occurrences are sinusoidal over the simulations. Each holds
265 positions outside of standard B-DNA conformations with the widening direction towards HTH2 having
266 a lifetime of 250-300 ps, maximum 400 ps, and the contraction towards HTH1 a maximum life of 80 ps,
267 and average of 40 ps. The widened groove is occupied by the Loop domain, and in contrast is contracted
268 in the x-ray structure, whereas the TATA region in published X-ray structures is not comparable due to
269 the base T13 flipping, the minor groove from T11-A14 has a 8.9 Å width in separate published x-ray
270 structures. These highlight the further differences between static x-ray DNA-protein interactions and
271 simulations at physiological conditions necessary for the study and development of DNA binding
272 proteins.

Effects of N-terminal Protein Processing

Tc3a processed N-terminal amino acids have a strong effect on DNA affinity. We conducted steered molecular dynamics to determine the binding affinity of the Tc3a bipartite domain, and the respective HTH1 and HTH2 individually, along with the connecting loop region. Total affinity was higher than expected from published work, where the average HTH has been shown to be around 80-120 k Joules/mol, although the total affinity was almost exactly the same for dozens of complete DNA binding transcription factors, around -400-440 kJoules/mol [47–49]. Our simulations showed the entire Bipartite of Tc3a Gibbs free energy change to be -416.17 k Joules, with the change in free energy presented in reverse of the binding process in Fig 3A. However analysis of the domains showed a normal, expected range for HTH2, -114.5 kJoules. Most of the energy could be attributed to the HTH1 domain and loop region, with a -155.7 and -148.2 kJoule/mol affinity respectively. This was higher for the HTH1 than expected, but was extremely high for a small 12 amino acid loop that in studied transposase only resides in minor grooves making primarily phosphate based interactions, Fig 3.

Fig 3. Free Energy Change. A] Binding affinity of the entire bipartite domain averaged from all pulled simulations, **B]** calculated mean affinity for HTH2, these did not vary between different N-terminal modifications, **C]** Binding affinity means for all modified N-terminus, blue, acetyl-Pro-Arg, green, NH2-Pro-Arg, red, removed Pro-Arg, NH2 terminal, black, NH2-Met-Ala-Gly, **D]** Binding affinity means for Loop region, residues Gly50-Ala60, standard deviation are given as +/- either side of mean for A-D].

To further dissect the interactions responsible for the higher affinity, we repeated the steered molecular dynamics simulations with changes to the N-terminal region. This resulted in 3 separate

297 experiments, removal of the acetyl group replaced with NH2-Pro, removal of the terminal residues Pro-
298 Arg with a truncated NH3, and replacement of the N-terminus with methionine, alanine and glycine as
299 the terminal residues. Higher DNA affinity can be attributed to the acetyl group and primary 2 N-
300 terminal amino acids. In all simulations, no significant difference was found for HTH2, as presented in
301 Fig 3B. Our most profound change was in removal of the N-terminal Pro-Arg, which resulted in a
302 binding affinity of -90.9 k Joules/mol for HTH1, while both the NH2-terminal-Pro and alternate Met-
303 Ala terminus had roughly equal affinities at the higher range of expected HTH motif affinities. Alone
304 this showed the acetyl group itself contributes 25-30 kJoules to the HTH1 binding process, higher than
305 can be attributed to any hydrogen bonding effects from a single oxygen, (Fig 3C).
306 Unexpectedly, we also found extreme effects on the calculated loop affinities, (Fig 3D). In contrast
307 to the HTH1 the NH2-Pro structure had a much lower affinity, -96 kJoules, even lower than removal of
308 the terminal residues. This was due to the NH2 and the proline preventing, or hindering some of the
309 interaction of Arg54. However, the altered N-terminus showed a higher than expected affinity of -129
310 kJoules. In simulations, the methionine and alanine are able to partially occupy the same spaces as the
311 natively processed N-terminus due to hydrophobic effects, without altering DNA structure, (Fig 4). This
312 allows Arg54 to make hydrogen bonds with DNA bases A43 and G15, however the methionine methyl
313 group prevents the Arg54 from entering deeper into the minor groove, also hindered by the lacking slight
314 widening found in the natural interaction. These hydrogen bonds are rigid, and differ from the Acetyl
315 processed conformation by having a constant hydrogen bond with the opposite strand, rather than one
316 side in the native protein. This further prevents the loop from occupying deeper into the groove,
317 especially Ala55 and Pro56 which are driven by hydrophobic forces from the solvent. None of the other
318 simulations flipped T13 out into the solvent, or any other bases in the ITR.
319

320 **Fig 4. Methionine-Alanine-Glycine N-terminus Bound to ITR DNA.** The modified N-terminus does
321 not move deep into the minor-groove, and Arg54 has different hydrogen bonding than the unmodified
322 terminus structure. DNA is shown as cartoon, select bases and amino acids shown as stick models. In
323 both, hydrogen, white, nitrogen, blue, oxygen, red. In amino acids, carbons, light blue, in DNA, carbons,
324 light purple. Hydrogen bonding from Arg54 is shown as dotted yellow lines.
325

326 # Inverted Repeat dynamic DNA structure

327

328 We further analyzed the changes in the ITR DNA from steered simulations, and unrestrained long
329 simulations to analyze the effects on the overall base pairing and architecture. In the binding process,
330 the majority of changes correlate to distances already determined from the longer simulations, (Fig 5A),
331 with the ITR stretch adopting perfect DNA architecture without protein. Base flipping generates a force
332 on the DNA equal to a right handed vector, moving in the 3' to 5' direction of the strand involved. This
333 causes a contraction of the minor groove, or the DNA as a whole if unrestricted. The bipartite in general
334 seems to cause the force to propagate in one direction, stabilized by the HTH2 end holding the DNA in
335 a classical conformation, and utilized by the HTH1 end to allow stronger affinity. A net effect is the
336 force moves in the direction towards the opposite ITR as found on the transposon, which alone or together
337 in the opposite direction with the second Tc3a would move the propagation of force into the DNA
338 encoding the transposase between the two end [50, 51].

339

340 **Fig 5. RMSF and Base Pair Densities for Extended Simulations. A]** ITR DNA shown as jelly role,
341 colored by motion normalized as B-factors from 0.1%-100%, white, 0% motion, thickness also indicates
342 greater motion scaled to the same B-factor. Red arrow, HTH1 binding site, green arrow, loop binding
343 site, blue arrow, HTH2 binding site, black arrow, flipped T13. **B]** Base pair C24-G32, shows an almost
344 normal non-moving base pair density as reference, **C]** the flipped T13-A43 base-pair, arrows indicate the
345 positions of phosphates, black, A43, red, T13, **D]** G15-C41, showing a transition to less movement as in

346 B], with slightly wider spacing, **E]** T11-A45, showing an extended spacing but still present centered base
347 pairing of the terminal bases. In all, density a percentage of total space occupied for all trajectory frames
348 in the x-y plane using the DNA central axis between base pair shown, including +/- 3.5 Å in the Z plane
349 to include the ribose and phosphate.
350

351 To further analyze the effects on individual base pairs, we conducted a density profile for each base
352 pair along the ITR over 50ns from 100ns simulations. Both bp 4-52 and 24-32 show profiles close to
353 what is expected geometrically in fluid simulations, with each base showing 80-90% of the space
354 occupied by the ribose ring in a fixed position, (Fig 5B and S1 Fig) [52, 53]. Base T13 remains in a
355 flipped solvent exposed conformation the entire simulation, (Fig 5C), with much less movement in the
356 position of the normally paired A43 base. Contraction of the minor groove in one direction causes a
357 lengthening of the base interactions from bases A12 to C10, which is normalized around base T8. The
358 majority of this lengthening is observed in bases T11 and A12, (Fig 5E and S1 Fig). In the opposite
359 direction, the effect seems to mainly be with base pair interactions themselves moving slightly off at an
360 angle, making the phosphate and ribose move more than the bases themselves. This dissipates by base
361 T19 to T20, with base C24 falling into normal expected DNA occupancy. An exception was base A16,
362 which showed low movement, however transitioned between an extended base pair, similar to bases T11
363 or C10, and a bent conformation like adjacent residues. This is mainly due to the only transient
364 interaction with Lys53 and Arg57, around 10% of simulation time, where the remaining bases have
365 strong consistent interactions. Results of these later interactions prevent any force from the flipped T13
366 from moving in the HTH2 direct along the ITR, as the bases are stabilized by amino acid interactions in
367 the deformed state.
368

369 **Bipartite Dynamics**

370

371 Applying RMSF to the bipartite in a bound and unbound state revealed the minor dynamics of the
372 protein domain. This showed results expected from simple HTH domain binding proteins, with HTH1
373 slightly more rigid in the bound conformation than HTH2, which contrasted the ITR fluctuations, (Fig
374 6). The Acetyl-Pro to residue Asp9 went from low to almost 90 percent fluctuation, meaning they are
375 moving freely in the solvent. This also translates into more springlike motion for the associated Helix
376 one of HTH1, with the end Ser8 to Glu11 forming a helix that constantly extends then recoils. This is
377 mirrored for the loop region from Tyr49 to Arg58, which is stabilized against the minor groove of the
378 ITR through hydrogen bonding, with Lys53 showing the most movement. In simulations Lys53 locks
379 over Arg54 to a phosphate oxygen only 40% of the time, further aiding the Arg maintain a one stranded
380 hydrogen bonding conformation.

381

382 **Fig 6. RMSF of Bipartite Domain. A]** Jelly role representation of the unbound and **B]** DNA bound form
383 of bipartite showing mean RMSF of all amino acids. B] shown as bound to ITR DNA, color, thickness
384 represented as 0.1%-100%, 0% white, motion as normalized B-factor according to color scale shown.
385 **C]** Normalized individual amino acids as bar graph scaled as in A,B], bars in numbering at midpoint
386 between bound, black and unbound, red. The last bar, residue 104. is the same for both conformations
387 (red not shown).

388

389 A remainder of motion changes are one of two, first the helices in both HTH become more compact,
390 slightly rolling hydrophobic residues inward when unbound, allowing residues such as HTH1 Lys19 to
391 Tyr42 and Lys31 to Thr10 or Asp9, to interact and hydrogen bond helix ends. This occurs with similarly
392 placed residues on HTH2, with Arg54 and Arg67, which are solvent exposed only in the bound form.
393 Secondly, residues such as Arg30, Arg54, Arg40 or Arg 94 move from a strong DNA base interaction to
394 being solvent exposed.

395

396 ## Dynamics of the Bipartite-ITR binding process

397

398 To determine the DNA base pair interactions giving rise to or affecting the higher affinity, we looked
399 at hydrogen bonding across all steered molecular dynamics simulations, averaged across all runs.
400 Initially we looked for only the Acetyl-Pro-Arg N-terminal interactions, and extended this to include
401 Gly4 for hydrogen bonding, (Fig 7). Between 3800-4800 ps, the hydrogen bonding becomes solely Arg3,
402 Gly4 peptide chain N and Ser5, and ends with a 2-400 ps stretch where only the Arg3 and Gly4 peptide
403 chain N make bonds longer than 1 ps. A number of transient bonds are made from the Arg3 functional
404 NH1, NH2 and NE hydrogens with T45, T13, T44 including the base oxygens or nitrogens, ribose O4',
405 and O1P or O2P lasting 1 ps or less, (Fig 7C). These interactions however are unable to account for the
406 energy involved based on hydrogen bonding alone [54].

407

408 **Fig 7. Total and Acetylated N-terminal Hydrogen Bonding. A]** Image of the N-terminal Acetyl-Pro-
409 Arg-Gly during the dissociation process, DNA cartoon, amino acids and T13, A45 as stick models,
410 oxygens, red, nitrogens, blue, carbon light blue, hydrogens white. Dotted yellow line shows the 3.5 Å
411 Acetyl-O to A45-O4' repulsive interaction. **B]** same as A], dots represent space filling model occupied
412 by outer atom electron shell, yellow dotted lines indicate hydrogen bonding between Gly4 and Arg3
413 mainchain nitrogen-hydrogen with the phosphate oxygen and O3' of A45. Both A and B are 180 degree
414 rotations of the orientation shown in Fig 2, A and B for comparison. **C]** Hydrogen bonding from pulled
415 simulations as a function of time, green, all hydrogen bonds from Acetylated-Pro-Arg bipartite, red, all
416 hydrogen bonds from NH2-Pro-Arg simulations, orange, only the hydrogen bonds from residues Acetyl-
417 Pro-Arg-Gly N-terminus. Black lines mark the means of each curve. Atom names are those used in the
418 Amber-ILDN force field from simulations.

419

420 We then looked at distances between base pairs, along with distances for the Acetyl group and
421 terminal amino acids. Our general finding was that the DNA stretch from A7 to A12 oscillates between
422 an A form and B form of DNA. When the HTH1 becomes disengaged, the Acetyl-Pro-Arg are the last
423 residues to leave. During these processes, the T13 begins to flip back into a normal conformation,
424 however the force causes A12 and T44 to lose their base pairing. During this process, the NH1, NH2
425 and NE of Arg3 form transient bonds pulling on A12 and T13 base ends. Mainchain nitrogens of Arg3
426 and Gly4 then interact with A45 ribose O3' and O1P, while the H3 of base T13 and N7 of the base ring
427 T45 form a long lasting stable hydrogen bond. This traps the acetyl group through hydrophobic forces,
428 and a strong repulsive effect between the oxygens of the group and T45 ribose O4', as the Acetyl-Pro
429 has no rotational movement between the Pro-Arg mainchain, (Fig 7A,B, and 8). This structure is stable
430 for 600-1200 ps in simulations. Interestingly the timing for this processes varied in pulled simulations,
431 and under bulk analysis conditions would be missed with averaged hydrogen bonding, RMSF or spatial
432 analysis techniques. We attribute this to the expected differences in starting points or positions of atoms,
433 as described for development of umbrella sampling methodologies requiring multiple starting points for
434 kinetic mean determination. The resulting free energy change for dislodging these interactions is 25-30
435 k Joules/mol and accounts for the increased affinity, which is determined from the bulk analytical
436 methods employed. None of the other simulations with modified N-terminus formed similar orientations
437 or structures, as the T13 was not unpaired or flipped out in any of these simulations.
438

439 **Fig 8. Distance Across Pulled Simulation Between Selected Atoms. A]** Base distance between atoms
440 N1 A12 and H3 T44, **B]** Base distance between atoms H3 T13 and N1 A43, **C]** Distance between atoms
441 O4' ribose ring A14 and O1P of T45, **D]** Composite distances between mainchain NH Arg3 and atoms
442 O4' ribose ring A14, black, O1P A45, red, and O4' ribose ring A45, green **E]** Composite distances
443 between atom NH mainchain Gly4 and atoms O4' ribose ring A14, black, O1P A45, red, and O4' ribose

ring A45, green **F]** distances between atom Acetyl O and O4' ribose ring A45, black, and atom N7 base ring A45 and ring H3 T13, red.

Proteomic analysis of selected N-terminal protein groups

Utilizing computational proteomics, we attempted to determine if there were any obvious correlations between the processed Acetyl-Pro-Arg N-terminus and DNA binding proteins. We also included dsRNA or RNA/DNA in the analysis, as a structural feature. Using a search algorithm, we extracted all proteins with the respective N-terminus from the Uniprot database, grouping this according to the datasets as presented from source. Using Pfam, Interpro and GO biochemical process identifiers, we then grouped all these into dsDNA and dsRNA binding proteins, (Table 2). Discrepancies exist between Tremble, which includes all sequenced data, and Swiss-protein which has heavily curated and accepted entries. At the organismal level, Plants and Achaea only showed major differences using both dataset types in combination. Bacteria were not included, as they lack acetylation of N-terminal residues, and Achaea only have similar processing of N-terminus in half of sequenced species [55].

Table 2. Tabulated Uniprot Parsed N-terminal Acetyl-Pro-Arg Proteins

Uniprot Dataset	Number Entries	DNA Associated	RNA-Associated
H.sapíens			
Tremble	358	37.70%	40.20%
Swiss-prot	129	69.80%	67.40%
Rodent			
Tremble	4546	50.70%	47.60%
Swiss-prot	229	60.30%	60.70%
Mammals			
Tremble	25416	41.80%	42.00%
Swiss-prot	74	40.50%	52.70%
Vertebrates			
Tremble	11858	39.30%	40.00%

Swiss-prot	198	60.10%	59.10%
Invertebrates			
Tremble	52472	27.70%	28.60%
Swiss-prot	89	61.80%	77.50%
Plants			
Tremble	48196	19.90%	23.80%
Swiss-prot	156	22.40%	28.20%
Fungi			
Tremble	109083	27.80%	32.70%
Swiss-prot	154	63.60%	74.00%
Archaea			
Tremble	23518	20.80%	29.30%
Swiss-prot	104	43.30%	76.90%
Virus			
Tremble	9827	36.40%	36.90%
Swiss-prot	325	63.70%	62.70%

Number of proteins, grouped according to Uniprot datasets found with N-terminal Acetyl-Pro-Arg. Percentages are all proteins found with GO function, Pfam, and Interpro composited for DNA or RNA interacting. Overlap exists between DNA and RNA interacting as a number of proteins are also dsDNA/RNA interacting or ribonuclear proteins that interact with dsDNA. A number of Tremble and a small portion of Swiss protein (~10-15% and 1-2% respectively) have unknown functions. Column 1, organism division, and Tremble or Swiss protein dataset, column 2, number of entries found, column 3, percent with dsDNA interaction identifiers, column 4, percent with dsRNA or ribonuclear protein identifiers.

To further refine and compare this data, we extracted data sets for H. sapiens only using additionally Met-Pro-His(MPH), Met-Pro-Glu(MPQ), Met-Ser-Arg(MSR), Met-Ala-Leu(MAL) and Met-Arg-Glu(MRQ) for all protein sequences, (Table 3). These last three were randomly chosen by using a simple random number generator for each sequence at positions 2 or 3 assigned a number from 1-20, while the Pro-His and Pro-Glu were chose as a similar, and opposite charge control. Then, data was refined to remove alternate spliced isoforms, and only curated proteins which have interactions already known. In some cases only a single isoform with an alternate methionine starting point contains the sequences listed, differing from another transcript with alternate starting 3 residues. Based on this, we determined all

100% known dsDNA, dsRNA and dsDNA/RNA proteins from the datasets. Statistically, the Acetyl-Pro-Arg was slightly higher than the others, 54.55%, with MPH, MPQ, MSR, MAL at around 30.00% each, and the MRQ which is not processed with methionine removal and acetylation at 7.92%. Enrichment analysis against the human genome showed Acetyl-Pro-Arg had 63 proteins out of 126 with a p-value <0.01 that were associated with dsDNA or dsRNA binding, 32 of which were transcription factors. However MRQ showed zero enrichment for any molecular function, and MAL showed only enrichment for cytokine and interferon receptor associations. Only the MSR, processed to acetyl-Ser-Arg in H. sapiens, showed 30.0% of proteins as enriched for direct transcription factors and DNA binding.

Table 3. Comparison of different H.sapiens N-terminal extracted proteins from Uniprot datasets.

N-terminus	Number-Annotated	Number dsDNA/ RNA binding	Percent	MOET *a
Ace-PR	112	61	54.55%	50% DNA transcription/ 8% DNA editing/ 23% RNA binding
Ace-PH	105	36	34.29%	50% Oxoreductase, 35% single nucleotide binding, 20% lipid oxidizing, 5% transcription factor
Ace-PQ	108	32	29.63%	23% α-actin associated
Ace-SR	150	45	30.00%	35.10% transcription factor
Ace-AL	344	73	21.22%	48.5% Cytokine receptor associated, 20% interferon receptor associated

MRQ	101	8	7.92%	0% enrichment

The N-terminus used is listed to the left. Only MRQ is not processed by human processing machinery into acetylated, methionine cleaved N-terminus. Column 2 is the number of proteins completely annotated with the respective N-terminus. Column 3 is the number of proteins that have direct dsDNA, dsRNA or dsDNA/RNA interactions listed in their functional identifier entries. Column 4 is the percentage shown to be ds nucleotide interacting, columns 2-4 processed through DAVID. Column 5 is the entire Uniprot H.sapiens Tremble and Swiss Protein datasets with all alternate spliced isoforms, redundant entries were remove, parsed through MOET to identify functional enrichment, having functional annotation including "probable", shown as percent of total included proteins, (342 ace-Pro-Arg min to 1117 ace-Ala-Leu max]), with p-values <0.001.

*a, a large number of alternate isoform entries, and "unknown function" in Tremble lack annotation. These were removed in the MOET analysis, however may effect statistics if and when they are accepted and annotated in future genome releases.

Discussion

Our data clearly shows the acetylated processed N-terminus plays an important role in the Tc3a transposition process. Here the effect is approximately 77 k Joules higher affinity compared with the NH2-Pro-Arg alone showing the difference cannot be attributed to hydrogen bonding, with only a single additional oxygen. The processes involved rather result from drastic differences in DNA structure induced by the different N-terminal structures highlighting the sharp effects of small molecular changes. Acetyl groups only have 6 atoms, with only 1 charged oxygen and in the Tc3a N-terminus only remove 2 hydrogens from the Pro. This resulting change causes a flipping of T13 in the TATA region of the Tc3a ITR, which is not mirrored in other N-terminus tested here, including the addition of an unprocessed methionine. Higher affinity is the result of this induced DNA conformation, and the effects it has on the DNA during the binding and dissociation processes of the Tc3a bipartite, as presented in Fig 7 and 8. A

517 novel stable DNA-protein intermediate is responsible for 25-30 k Joule differences observed from all
518 tested N-terminus in the study, which occurs for an extended period in molecular time frames between
519 600-1200 ps. This is enough time for the HTH1 domain of Tc3a to dissociate and re-associate in natural
520 conditions, without an applied force. Additional higher affinity is achieved by the changes in the minor
521 groove, particularly with the loop region binding, also induced by the acetyl group.

522 Base flipping has been observed in other transposon families, outside of mariners, and has been
523 proposed to be a common mechanism in many DNA binding protein's function [56]. In Tc3a or similar
524 related transposase, the process most likely plays a role in the initial ITR cleavage process, which has
525 several competing proposed mechanisms [17, 18, 57]. Currently these include supercoiling to bring
526 single transposase together, dimerization at one or both ends, and a multimerization of 4 single mariners.
527 Torsion provided by the single base flipping at either ITR would produce a supercoiling of dsDNA,
528 where the 1200 base pairs between would wind around itself bringing the two ends together. Of the
529 debated mechanisms with supercoiling, one hypothesis includes a single nick at either end and bending
530 of the ITR cause a similar process, however base flipping has never been observed in Tc3a or mariners.
531 Much of these differences in theoretical processes may be an artifact of the expression systems used,
532 with all x-ray and NMR structures published utilizing E. coli which lacks the acetylation processing. It
533 has also been mentioned in literature that the HTH1, or bipartite domains of mariners have lower than
534 expected affinities in DNA binding wet laboratory experiments, which utilize the same expression
535 systems and are routinely used in most wet laboratory experiments to develop all process models.

536 Our deduced process however may be Tc3a specific, with only 4% of known mariner transposons
537 having an N-terminal acetylated Pro-Arg, although this number increases slightly if Mos1, piggybac, or
538 related mariner splice variants are included to 10%. Also a number of transgenic mariner evolved fusion
539 proteins such as SETMAR in humans and Papio anubis retain the N-terminus [58, 59]. Interestingly

540 Phytophthora and Planariidae have a larger portion of identified mariner transposase with this Acetyl-
541 Pro-Arg N-terminus, grouped into Tc3 like or Tc1 like. Still, without further sub-grouping transposase,
542 mariner transposase must have alternate mechanisms for the same process if this base flipping is intrinsic
543 to the family as a whole. It has been shown in Tn5 and Tn10 transposase, which utilize different
544 mechanisms including hairpin formations, base flipping 3-4 bases from the ITR cleavage site is mediated
545 by forcing a Methionine between the bases, aided by the bound transposase. Similar processes to the one
546 described in our work may be possible with alternative sequences, which would include 8000 novel
547 variants if all 3 primary amino acids are included. Ultimately, our work indicates that N-terminal
548 processing needs to be taken into account in the study of DNA binding proteins in general.

549 More theoretical, we attempted attempted to find a significant correlation with the Acetyl-Pro-Arg N-
550 terminus and dsDNA, dsRNA or dsDNA/RNA binding proteins. Our data indicates a slightly higher
551 correlation for the N-terminal residues analyzed, however the increase of only 20-25% more from others
552 indicates this may be in a more specific protein domain or family aspect. In simple enrichment analysis,
553 the Acetyl-Pro-Arg did show high enrichment over other tested N-terminal sets, however when looking
554 at the protein sets individually all contained DNA binding proteins. Additionally enrichment analysis
555 routinely left out proteins listed as direct dsDNA or dsRNA binding in their descriptions. In a number
556 of mammalian mariner evolved fusion proteins, the HTH1 and HTH2 are on the C-terminus, while the
557 Acetyl-Pro-Arg still make up the N-terminal domain, and are in direct contact with the DNA, suggesting
558 this still may have some role. We were however hindered from direct structural analysis as only 8
559 structures from the 487 human proteins, including splice variants identified, (Table 3), had structures
560 complete enough to analyzed. From these 3 showed the processed end formed a perfect β-sheet that was
561 twisted in the x-ray structures lacking processed N-terminus, 3 were ribosomal structural proteins
562 embedding the N-terminus in dsRNA, 1 was a ribonuclear protein with dsDNA/RNA, and 1 a

transcription factor superposable with Tc3a. This sample size is to small to make direct conclusions or provide statistic on function, making our proteomics analysis only suggestive for common functional categorizations.

Our work here shows the mechanism is an intrinsic part of the Tc3a transposase transposition process, conferring higher affinity and altering the DNA structure. For Tc3a specifically, the flipped base at either end of the transposon would provide enough force to cause the DNA between the sites to supercoil. This supports a model outlined by other groups (S2 Fig), where single transposase bind at either end and are brought together to form a dimer from DNA torsion alone. A model similar to this has been proposed as a general mechanism used to bring together distant transcription factors or DNA binding proteins in larger complexes [45,46,51]. Still, this does not address the order of DNA strand cleavage, which still contains models of single strand or double strand nicking also playing a role of bringing two transposase together. However, it does indicate that multimerization is not necessary for the initial inverted terminal repeat torsions necessary for the overall process. Thus more complex 4 transposase multimeric systems are most likely not involved in mariner ITR cleavage. We propose a more simplistic system based on the DNA binding from processed terminal residues and induced torsions in the ITR as a major factor driving dimerization.

ACKNOWLEDGEMENTS

Use of grid computers to conduct initial and subsequent simulations was made possible through. Laboratório Nacional de Computação Científica - Av. Getúlio Vargas, 333, Quitandinha, Petrópolis, RJ, Brasil, CEP 25651-075 and CENAPAD-SP

Rua Saturnino de Brito, 45, Cidade Universitária,Campinas, SP, Brasil, CEP 13083-889.

References

1. Mc CB. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950 Jun;36[6].344–55.

2. Mc CB. Maize genetics. Year B Carnegie Inst Wash. 1946;45.176–86.

607 3. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev
608 Genet. 2007;41.331–68.

609 4. Plasterk RHA, van Luenen H. Transposons. In. Riddle DL, Blumenthal T, Meyer BJ, Priess JR,
610 editors. C elegans II. Cold Spring Harbor [NY]. Cold Spring Harbor Laboratory Press Copyright
611 © 1997, Cold Spring Harbor Laboratory Press.; 1997.

612 5. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet.
613 2008 May;9[5].397–405.

614 6. Wang S, Diaby M, Puzakov M, Ullah N, Wang Y, Danley P, et al. Divergent evolution profiles of
615 DD37D and DD39D families of Tc1/mariner transposons in eukaryotes. Mol Phylogenet Evol.
616 2021 Aug;161.107143.

617 7. Gao B, Wang Y, Diaby M, Zong W, Shen D, Wang S, et al. Evolution of pogo, a separate
618 superfamily of IS630-Tc1-mariner transposons, revealing recurrent domestication events in
619 vertebrates. Mob DNA. 2020;11.25.

620 8. Wang W, Swevers L, Iatrou K. Mariner [Mos1] transposase and genomic integration of foreign
621 gene sequences in Bombyx mori cells. Insect Mol Biol. 2000 Apr;9[2].145–55.

622 9. Yokoo M, Fujita R, Nakajima Y, Yoshimizu M, Kasai H, Asano S, et al. Mos1 transposon-based
623 transformation of fish cell lines using baculoviral vectors. Biochem Biophys Res Commun.
624 2013 Sep 13;439[1].18–22.

625 10. Woodard LE, Wilson MH. piggyBac-ing models and new therapeutic strategies. Trends
626 Biotechnol. 2015 Sep;33[9].525–33.

627 11. Kim SW, Lee JH, Han JS, Shin SP, Park TS. piggyBac Transposition and the Expression of
628 Human Cystatin C in Transgenic Chickens. Animals [Basel]. 2021 May 26;11[6].

629 12. Plasterk RH. The Tc1/mariner transposon family. Curr Top Microbiol Immunol.
630 1996;204.125–43.

631 13. Wells JN, Feschotte C. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet.
632 2020 Nov 23;54.539–61.

633 14. Pace JK, and Feschotte C. The evolutionary history of human DNA transposons. evidence for
634 intense activity in the primate lineage. Genome Res. 2007 Apr;17[4].422–32.

635 15. Craig NL. Target site selection in transposition. Annu Rev Biochem. 1997;66.437–74.

636 16. Morris ER, Grey H, McKenzie G, Jones AC, Richardson JM. A bend, flip and trap mechanism for
637 transposon integration. Sherratt D, editor. eLife. 2016 May;5.e15537.

638 17. Ochmann MT, Ivics Z. Jumping Ahead with Sleeping Beauty. Mechanistic Insights into Cut-
639 and-Paste Transposition. Viruses. 2021;13[1].

640 18. Claeys Bouuaert C, Chalmers R. A single active site in the mariner transposase cleaves DNA
641 strands of opposite polarity. Nucleic Acids Research. 2017 Nov 16;45[20].11467–78.

642 19. Galbraith JD, Ivancevic AM, Qu Z, Adelson DL. Detecting Horizontal Transfer of Transposons.
643 Methods Mol Biol. 2023;2607.45–62.

644 20. Palazzo A, Escuder E, D'Addabbo P, Lovero D, Marsano RM. A genomic survey of Tc1-mariner
645 transposons in nematodes suggests extensive horizontal transposon transfer events. Mol
646 Phylogenet Evol. 2021 May;158.107090.

647 21. Dupeyron M, Baril T, Bass C, Hayward A. Phylogenetic analysis of the Tc1/mariner
648 superfamily reveals the unexplored diversity of pogo-like elements. Mobile DNA. 2020 Jun
649 29;11[1].21.

650 22. Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. Genes
651 Genet Syst. 2020 Jan 30;94[6].233–52.

652 23. Watkins S, van Pouderoyen G, Sixma TK. Structural analysis of the bipartite DNA-binding
653 domain of Tc3 transposase bound to transposon DNA. Nucleic Acids Res. 2004;32[14].4306–
654 12.

655 24. Wingfield PT. N-Terminal Methionine Processing. Curr Protoc Protein Sci. 2017 Apr
656 3;88.6.14.1-6.14.3.

657 25. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS. fast,
658 flexible, and free. J Comput Chem. 2005 Dec;26[16].1701–18.

659 26. Darden T, Perera L, Li L, Pedersen L. New tricks for modelers from the crystallography toolkit.
660 the particle mesh Ewald algorithm and its use in nucleic acid simulations. Structure. 1999
661 Mar 15;7[3].R55-60.

662 27. Humphrey W, Dalke A, Schulten K. VMD. visual molecular dynamics. J Mol Graph. 1996
663 Feb;14[1].33–8, 27–8.

664 28. Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, et al. Improvements to the APBS
665 biomolecular solvation software suite. Protein Sci. 2018 Jan;27[1].112–28.

666 29. Lill MA, Danielson ML. Computer-aided drug design platform using PyMOL. J Comput Aided
667 Mol Des. 2011 Jan;25[1].13–9.

668 30. Makarewicz T, Kaźmierkiewicz R. Molecular dynamics simulation by GROMACS using GUI
669 plugin for PyMOL. J Chem Inf Model. 2013 May 24;53[5].1229–34.

670 31. Hub JS, de Groot BL, van der Spoel D. g_wham.A Free Weighted Histogram Analysis
671 Implementation Including Robust Error and Autocorrelation Estimates. Journal of chemical
672 theory and computation. 2010;6[12].3713–20.

673 32. Lemkul JA. Introductory Tutorials for Simulating Protein Dynamics with GROMACS. J Phys
674 Chem B. 2024 Oct 3;128[39].9418–35.

675 33. Elber R, Ruymgaart AP, Hess B. SHAKE parallelization. Eur Phys J Spec Top. 2011 Nov
676 1;200[1].211–23.

677 34. Luzar A, Chandler D. Effect of environment on hydrogen bond dynamics in liquid water. Phys
678 Rev Lett. 1996 Feb 5;76[6].928–31.

679 35. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-
680 chain torsion potentials for the Amber ff99SB protein force field. Proteins. 2010
681 Jun;78[8].1950–8.

682 36. Briones R, Blau C, Kutzner C, de Groot BL, Aponte-Santamaría C. GROmaps. A GROMACS-
683 Based Toolset to Analyze Density Maps Derived from Molecular Dynamics Simulations.
684 Biophys J. 2019 Jan 8;116[1].4–11.

685 37. The UniProt Consortium. UniProt. the Universal Protein Knowledgebase in 2025. Nucleic
686 Acids Research. 2024 Nov 18;53[D1].D609–17.

687 38. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo. a sequence logo generator. Genome
688 Res. 2004 Jun;14[6].1188–90.

689 39. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. The Pfam
690 protein families database. embracing AI/ML. Nucleic Acids Res. 2024 Nov;53[D1].D523–34.

691 40. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, et al. InterPro. the
692 protein sequence classification resource in 2025. Nucleic Acids Research. 2024 Nov
693 20;53[D1].D44–456.

694 41. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER. Making
695 genome-scale phylogenetics accessible to all. Protein Science. 2022 Jan 1;31[1].8–22.

696 42. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID. a web server for
697 functional enrichment analysis and functional annotation of gene lists [2021 update]. Nucleic
698 Acids Res. 2022 Jul 5;50[W1].W216-w221.

699 43. Vedi M, Nalabolu HS, Lin CW, Hoffman MJ, Smith JR, Brodie K, et al. MOET. a web-based gene
700 set enrichment tool at the Rat Genome Database for multiontology and multispecies analyses.
701 Genetics. 2022 Apr 4;220[4].

702 44. van Pouderoyen G, Ketting RF, Perrakis A, Plasterk RH, Sixma TK. Crystal structure of the
703 specific DNA-binding domain of Tc3 transposase of C.elegans in complex with transposon
704 DNA. Embo j. 1997 Oct 1;16[19].6044–54.

705 45. Teves SS, Henikoff S. DNA torsion as a feedback mediator of transcription and chromatin
706 dynamics. Nucleus. 2014 May;5[3].211–8.

707 46. Reymer A, Zakrzewska K, Lavery R. Sequence-dependent response of DNA to torsional
708 stress. a potential biological regulation mechanism. Nucleic Acids Res. 2018 Feb
709 28;46[4].1684–94.

710 47. Jung C, Bandilla P, von Reutern M, Schnepf M, Rieder S, Unnerstall U, et al. True equilibrium
711 measurement of transcription factor-DNA binding affinities using automated polarization
712 microscopy. Nat Commun. 2018 Apr 23;9[1].1605.

713 48. Etheve L, Martin J, Lavery R. Protein-DNA interfaces. a molecular dynamics analysis of time-
714 dependent recognition processes for three transcription factors. Nucleic Acids Res. 2016 Nov
715 16;44[20].9990–10002.

716 49. Yesudhas D, Anwar MA, Panneerselvam S, Kim HK, Choi S. Evaluation of Sox2 binding
717 affinities for distinct DNA patterns using steered molecular dynamics simulation. FEBS Open
718 Bio. 2017 Nov;7[11].1750–67.

719 50. Randall GL, Zechiedrich L, Pettitt BM. In the absence of writhe, DNA relieves torsional stress
720 with localized, sequence-dependent structural failure to preserve B-form. Nucleic Acids Res.
721 2009 Sep;37[16].5568–77.

722 51. van den Broek B, Lomholt MA, Kalisch SMJ, Metzler R, Wuite GJL. How DNA coiling enhances
723 target localization by proteins. Proc Natl Acad Sci U S A. 2008 Oct 14;105[41].15738–42.

724 52. Arnott S. The geometry of nucleic acids. Prog Biophys Mol Biol. 1970;21.265–319.

725 53. Stofer E, Lavery R. Measuring the geometry of DNA grooves. Biopolymers. 1994
726 Mar;34[3].337–46.

727 54. Lide DR. CRC Handbook of Chemistry and Physics, 89th Edition [Internet]. Taylor & Francis;
728 2008. Available from. https.//books.google.com.br/books?id=KACWPwAACAAJ

729 55. Schulze S, Adams Z, Cerletti M, De Castro R, Ferreira-Cerca S, Fufezan C, et al. The Archaeal
730 Proteome Project advances knowledge about archaeal cell biology through comprehensive
731 proteomics. Nature Communications. 2020 Jun 19;11[1].3145.

732 56. Bischerour J, Chalmers R. Base flipping in tn10 transposition. an active flip and capture
733 mechanism. PLoS One. 2009 Jul 10;4[7].e6201.

734 57. Richardson JM, Dawson A, O'Hagan N, Taylor P, Finnegan DJ, Walkinshaw MD. Mechanism of
735 Mos1 transposition. insights from structural analysis. Embo j. 2006 Mar 22;25[6].1324–34.

736 58. Lié O, Renault S, Augé-Gouillou C. SETMAR, a case of primate co-opted genes. towards new
737 perspectives. Mob DNA. 2022 Apr 8;13[1].9.

59. Tellier M, Chalmers R. Human SETMAR is a DNA sequence-specific histone-methylase with a
broad effect on the transcriptome. Nucleic Acids Res. 2019 Jan 10;47[1].122–33.

60. Henri G. A. M. van Luenen, Sean D. Colloms, Ronald H. A. Plasterk. The Mechanism of Transposition of Tc3 in C. elegans. Cell. 1994 Oct 21;79.293-301.

61. David J. Lampe, Mair E.A. Churchill, Hugh M.Robertson. A purified mariner transposase is sufficient to mediate transposition in vitro. EMBO J. 1996, 15[19].5470-5479

Supplementary Material

S1 Fig. Occupied Density Maps of Base Pairs During Extended Simulations. Base pairs are listed above density plots, density is scaled as 0.1%-100% occupancy according to color scales on the right of each representation, white, 0% occupancy. Base pair 4-52 represents a low moving classical base pair as reference. Base pair 16-40 is an outlier. All represent the total occupancy from all frames, minus equilibration, taken from the DNA central axis between specific base pair as reference in the x-y direction, and +/- 3.5 Å in the Z direction.

S2 Fig. Example published proposed models of Tc3a Transposition Process. 1) A) A single transposase binds the ITR of the transposon at either end. **B)** Distortion on the DNA at either end causes the region between ITR to super-coil, bringing the two transposase together, allowing for the cleavage of both strands at the ITR end. **C)** The dimerized complex leaves the excision site, which is repaired by DNA ligase leaving a TA[xx]TA duplication*, **D)** remaining dimerized the complex then scans DNA until a TA site is recognized, inserts the transposon complete with the ITR at either end, and the two proteins dissociate. C and D remain the same for all models, except the coiling or looped DNA are both proposed in different models to make up the complex until insertion. **2) A)** Dimerized transposase binds either end of the transposon at the ITR, **B)** These cause torsion that bring the 4 transposase together, which then are able to cleave both strands of DNA, one of each original dimer leaves with the excision site, while the other two leave carrying the transposon. **3) A)** Two transposase bind one ITR of the

transposon **B)** these dimerize with unbound transposase forming a 4 protein complex excising one end of the transposon, and coil the DNA until the second ITR is found, and the entire transposon is cleaved at either ITR. This leaves the two excised ends distant from each other. Green ovals, transposase, red lines, ITR, blue lines, dsDNA, in **3B**, half red line a TA overhang.

***** The excision site at a lower frequency has a TAX_NTA marker, which can be 0-4 bases randomly inserted found adjacent to the proposed ITR cleavage site. Duplication refers to the doubled TA, the original insertion site between a T/A, known as the excision footprint. Variations exist on these 3 main proposed mechanisms, with more than a dozen proposed similar models having slight variations. [refs. 4,12,18,60,61 and other published works]

Complete equilibrated models used are deposited on the ModelArchive at https://www.modelarchive.org/10.5452/ma-5m69p

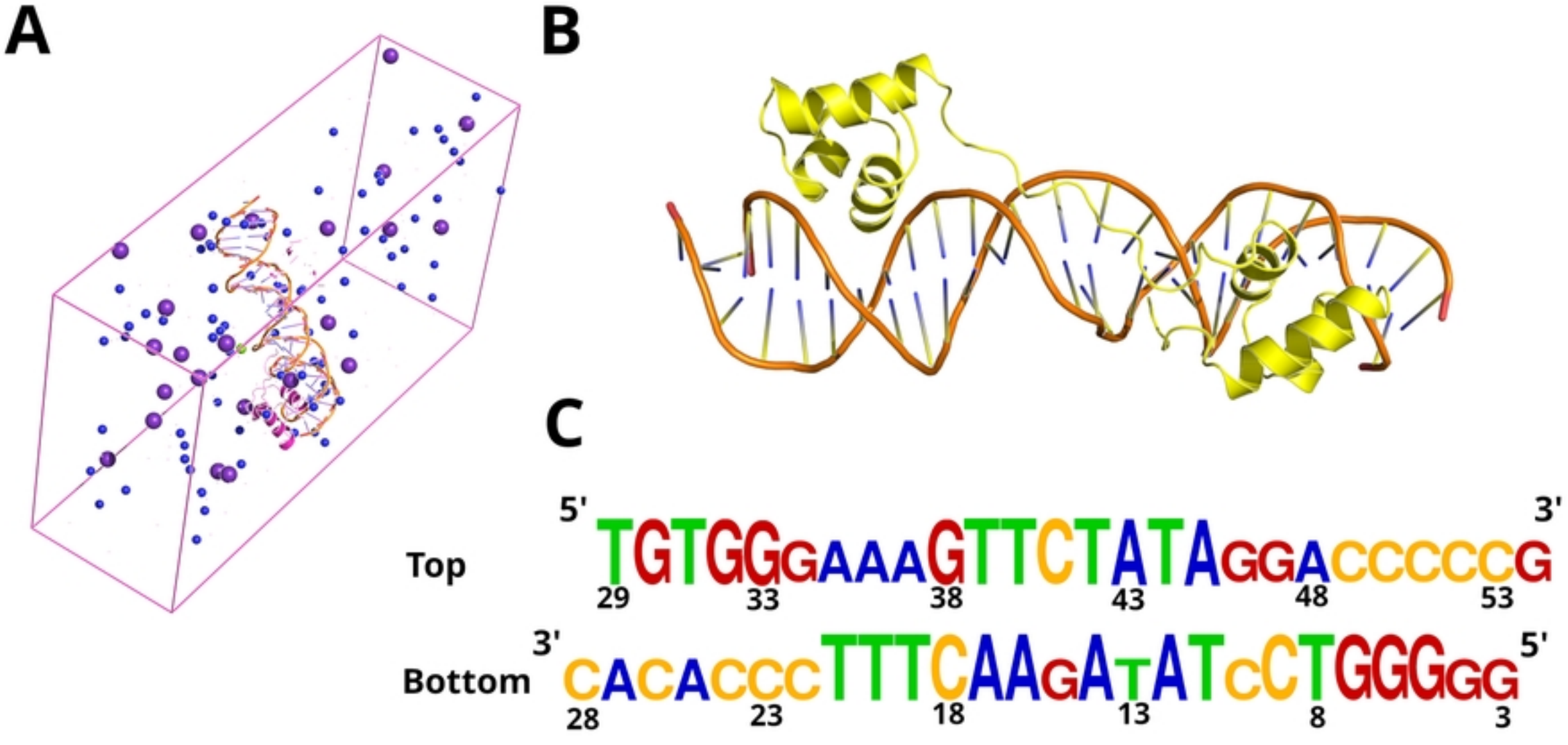
A
B
C
5'
3'
Top
TGTGGGAAAGTTCTATAGGACCCCCCG
29
33
38
43
48
53
3'
5'
Bottom
CACACCCTTTCAAGATATCCTGGGGGG
28
23
18
13
8
3

Figure 1

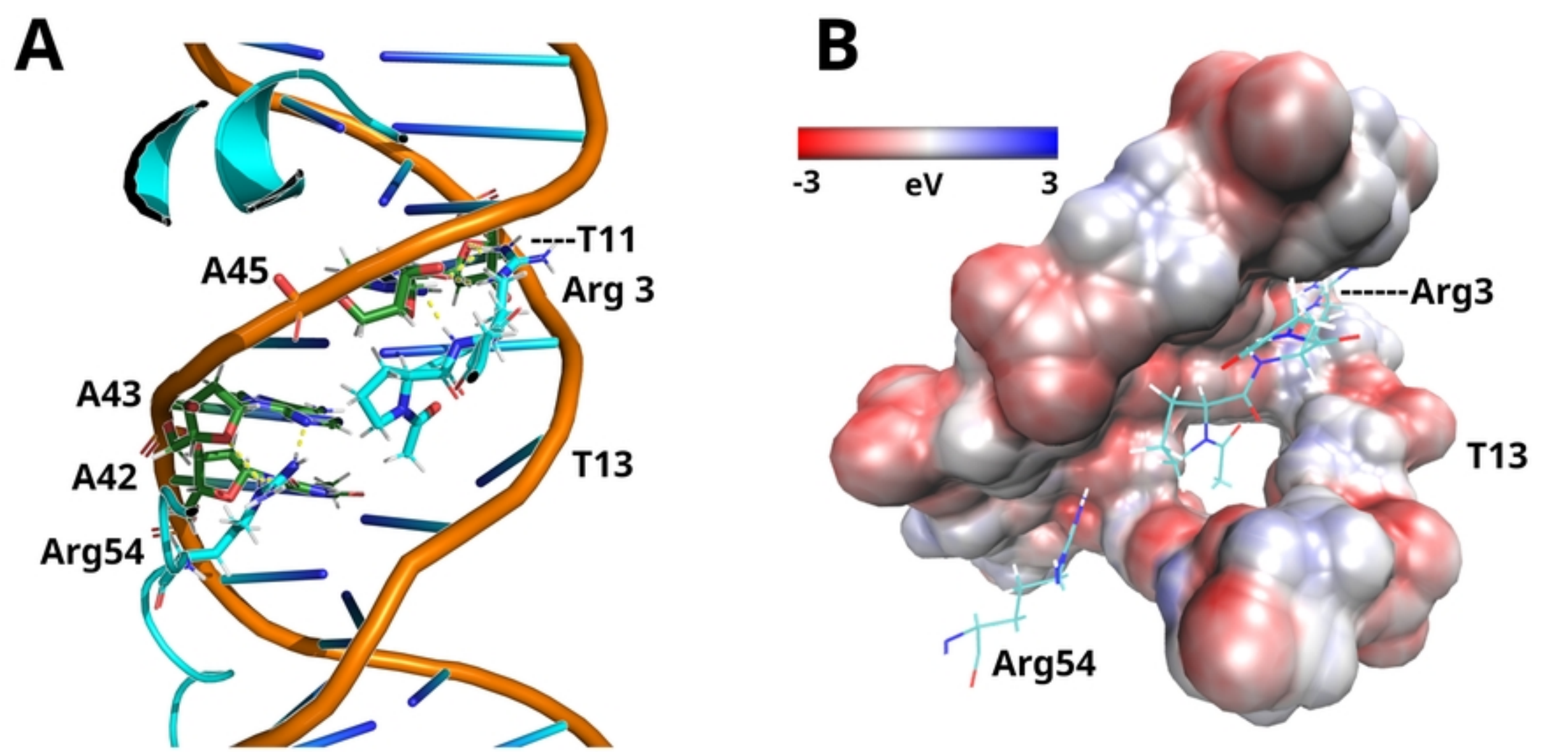

Figure 2

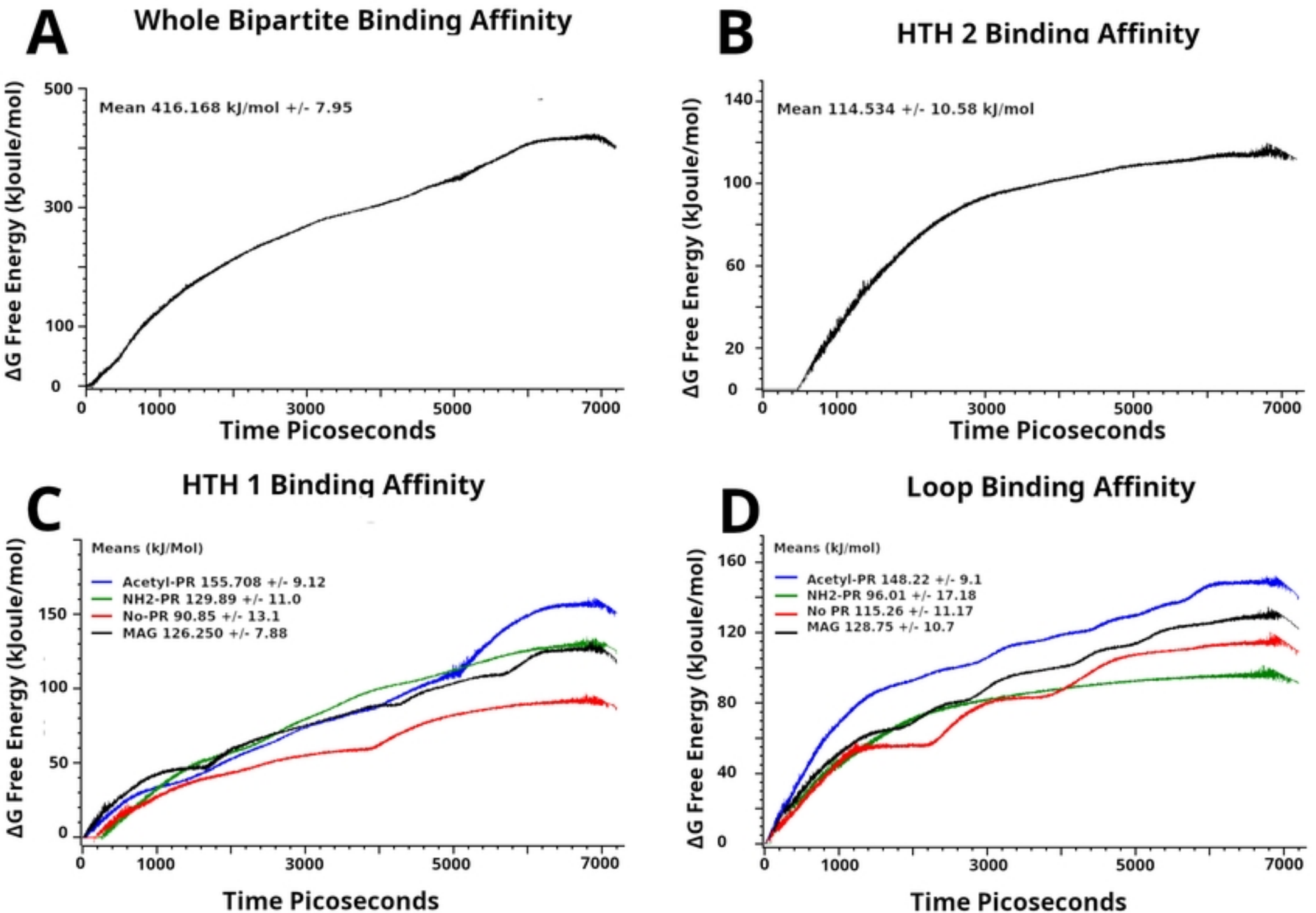
A
Whole Bipartite Binding Affinity
Mean 416.168 kJ/mol +/- 7.95
ΔG Free Energy (kJoule/mol)
Time Picoseconds
B
HTH 2 Binding Affinity
Mean 114.534 +/- 10.58 kJ/mol
ΔG Free Energy (kJoule/mol)
Time Picoseconds
C
HTH 1 Binding Affinity
Means (kJ/Mol)
Acetyl-PR 155.708 +/- 9.12
NH2-PR 129.89 +/- 11.0
No-PR 90.85 +/- 13.1
MAG 126.250 +/- 7.88
ΔG Free Energy (kJoule/mol)
Time Picoseconds
D
Loop Binding Affinity
Means (kJ/mol)
Acetyl-PR 148.22 +/- 9.1
NH2-PR 96.01 +/- 17.18
No PR 115.26 +/- 11.17
MAG 128.75 +/- 10.7
ΔG Free Energy (kJoule/mol)
Time Picoseconds

Figure 3

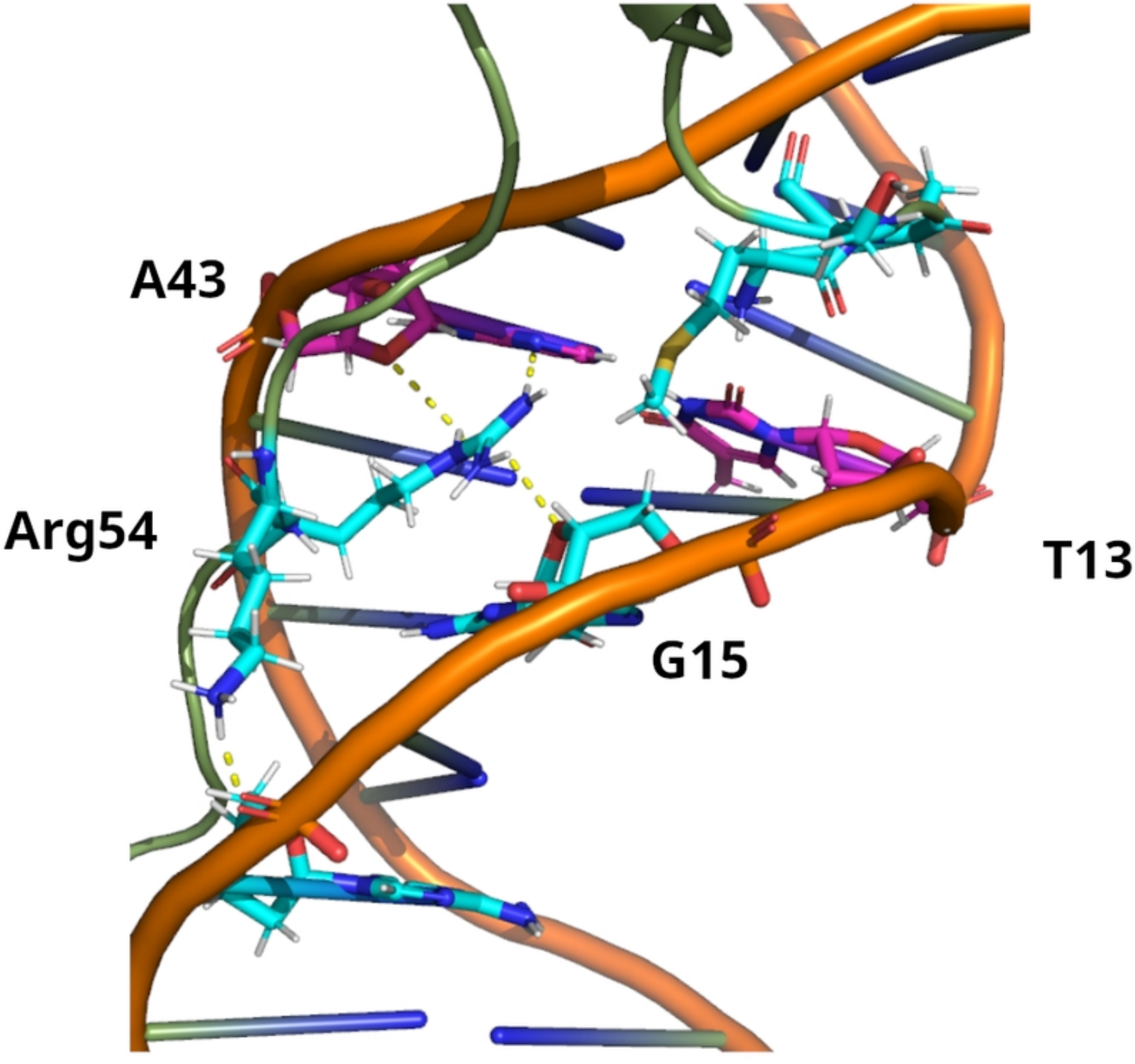

Figure 4

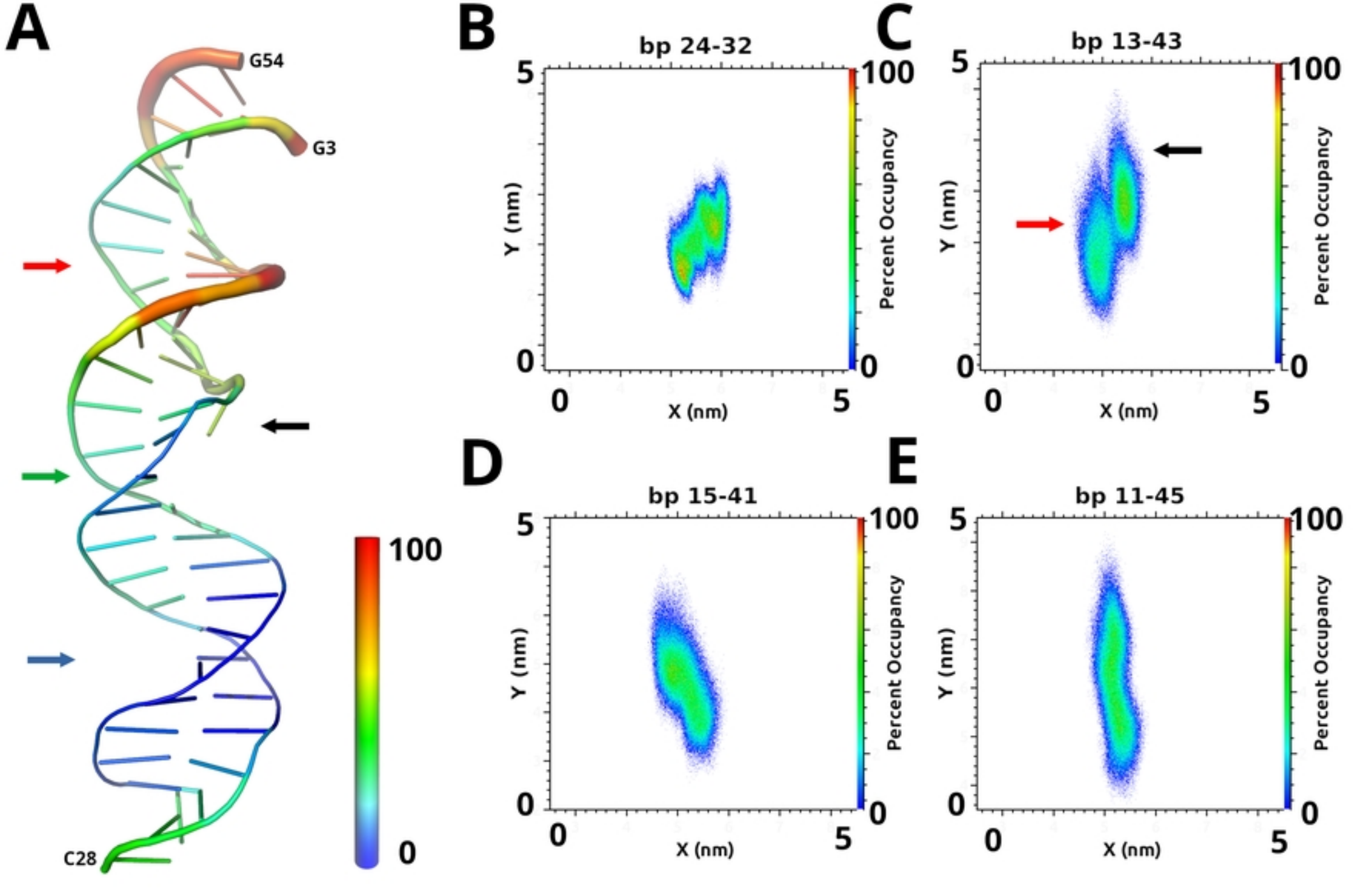
A
G54
G3
C28
100
0
B
bp 24-32
C
bp 13-43
D
bp 15-41
E
bp 11-45
Y (nm)
X (nm)
Percent Occupancy
0
5
100

Figure 5

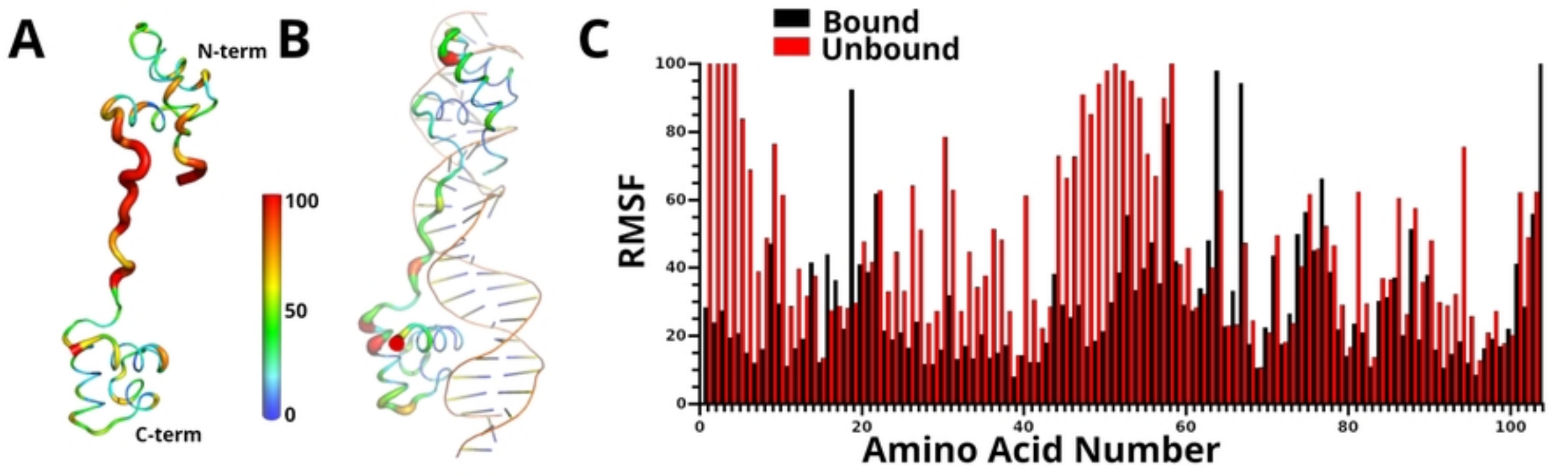
A
N-term
C-term
100
50
0
B
C
Bound
Unbound
RMSF
Amino Acid Number
0
20
40
60
80
100

Figure 6

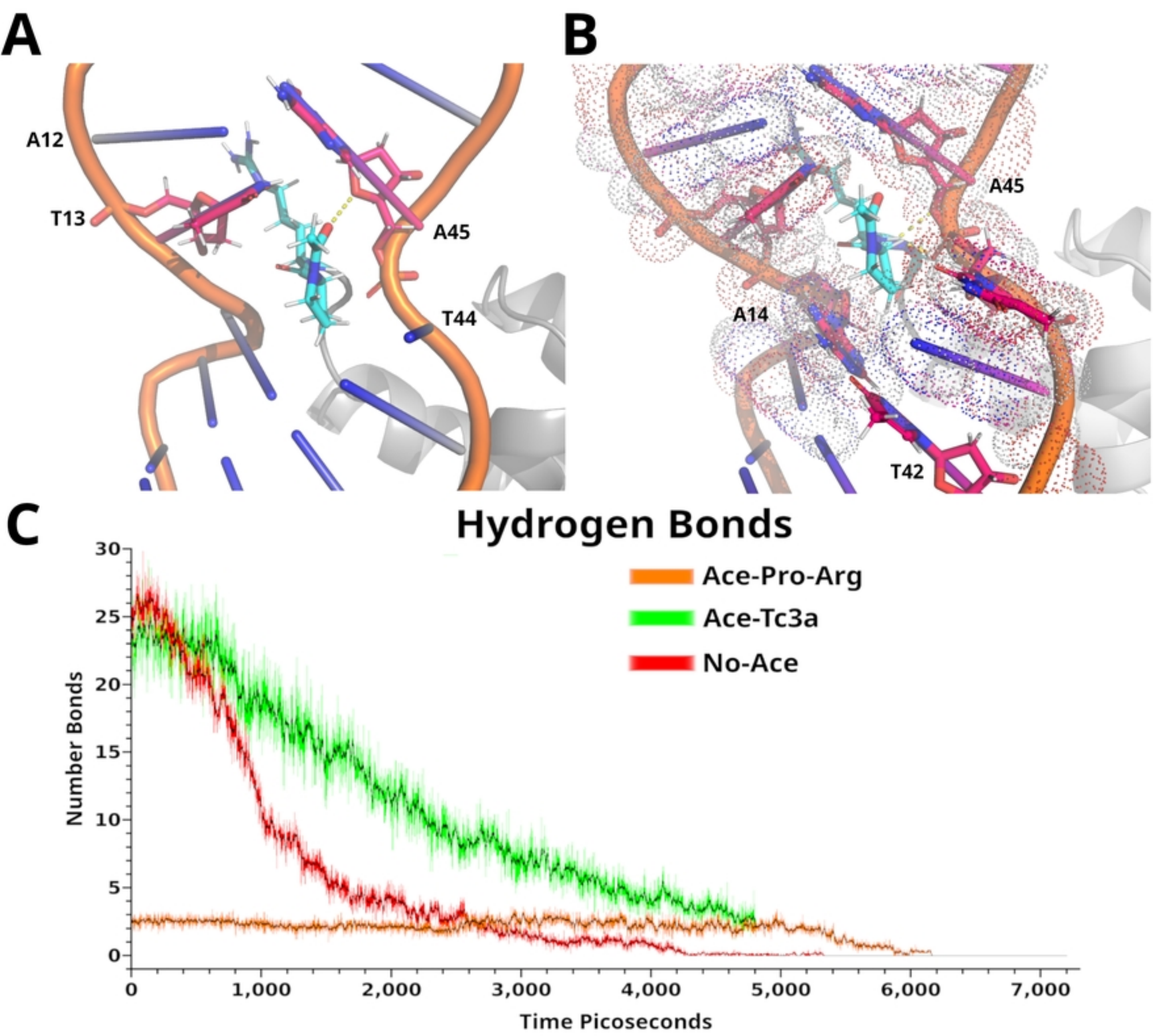
A
A12
T13
A45
T44
B
A45
A14
T42
C
Hydrogen Bonds
Ace-Pro-Arg
Ace-Tc3a
No-Ace
Number Bonds
0
5
10
15
20
25
30
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
Time Picoseconds

Figure 7

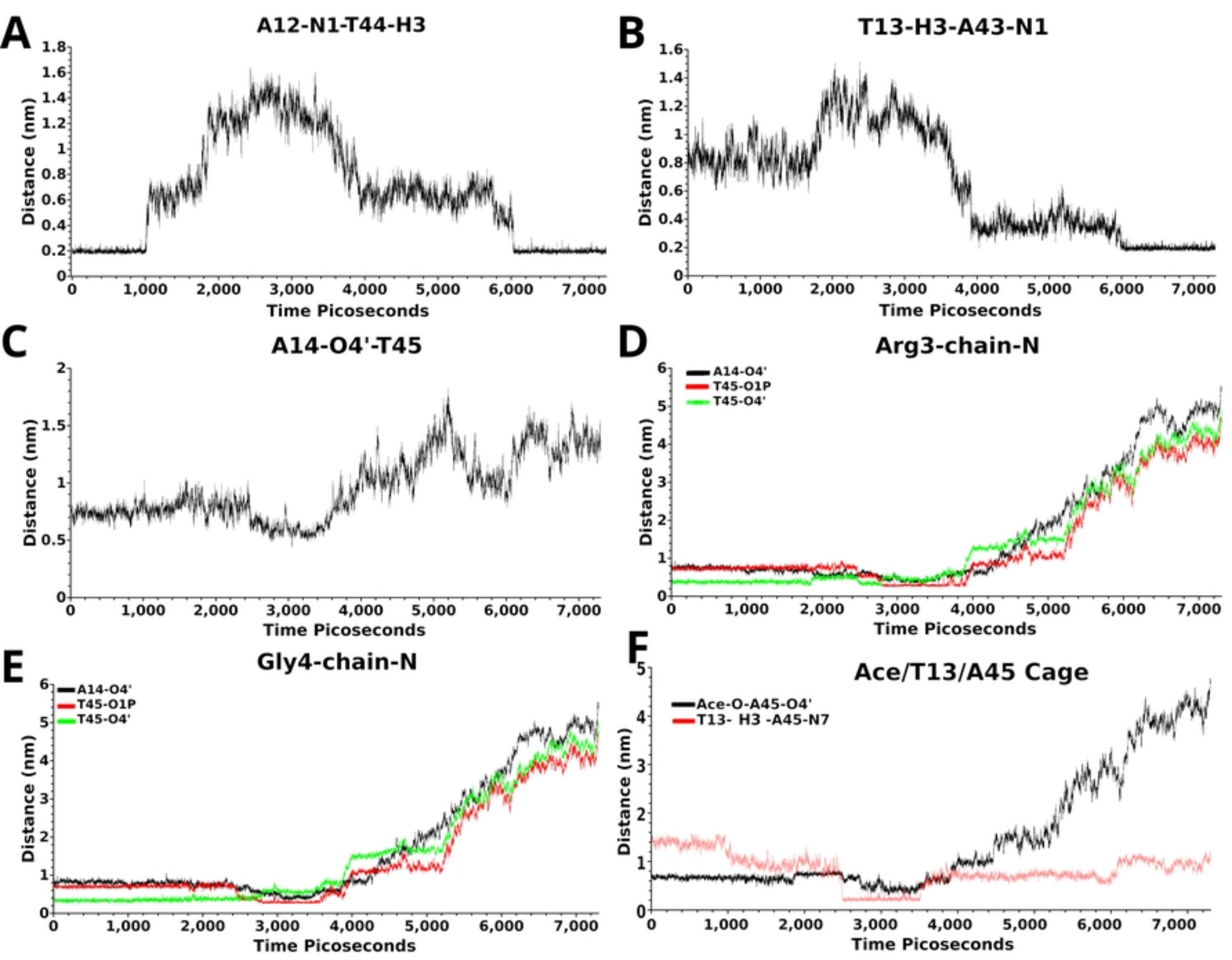
A
A12-N1-T44-H3
Distance (nm)
Time Picoseconds
B
T13-H3-A43-N1
Distance (nm)
Time Picoseconds
C
A14-O4'-T45
Distance (nm)
Time Picoseconds
D
Arg3-chain-N
A14-O4'
T45-O1P
T45-O4'
Distance (nm)
Time Picoseconds
E
Gly4-chain-N
A14-O4'
T45-O1P
T45-O4'
Distance (nm)
Time Picoseconds
F
Ace/T13/A45 Cage
Ace-O-A45-O4'
T13- H3 -A45-N7
Distance (nm)
Time Picoseconds

Figure 8