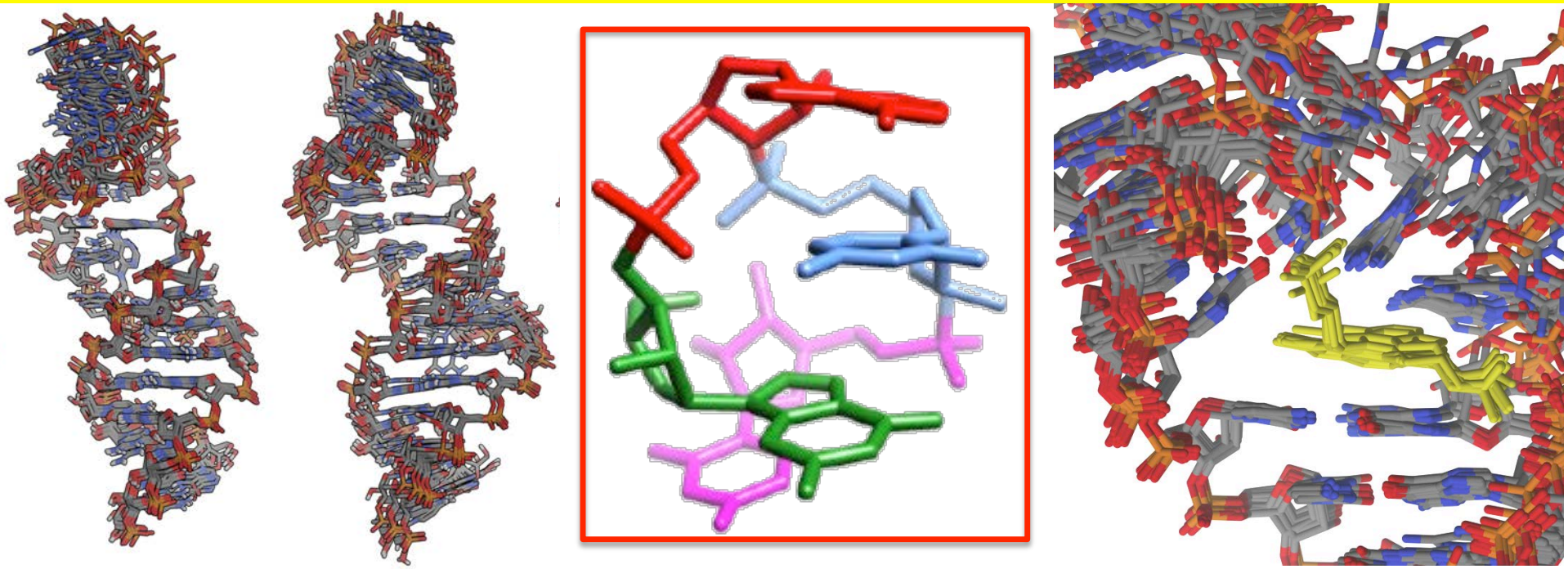
Success and challenge in modeling nucleic acid structure and dynamics …beware of too much sampling?



Thomas E. Cheatham III tec3@utah.edu
Professor, Dept. of Medicinal Chem., College of Pharmacy
Director, Research Computing and CHPC,
University Information Technology, University of Utah

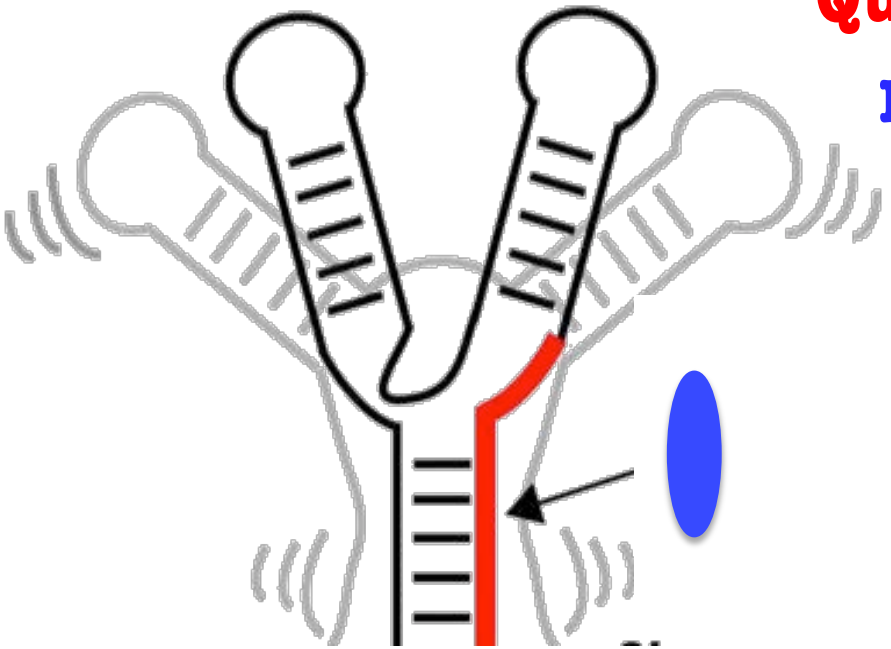
Ultimate goal: Use simulation to model RNA folding and ligand-induced conformational exchange / capture

This requires: Accurate and fast simulation methods

Validated RNA, protein, water, ion, and ligand "force fields"

"good" high-resolution experiments to assess results

dynamics and complete sampling: (convergence, reproducibility)



Question: Is the movement real or artifact?

assessment

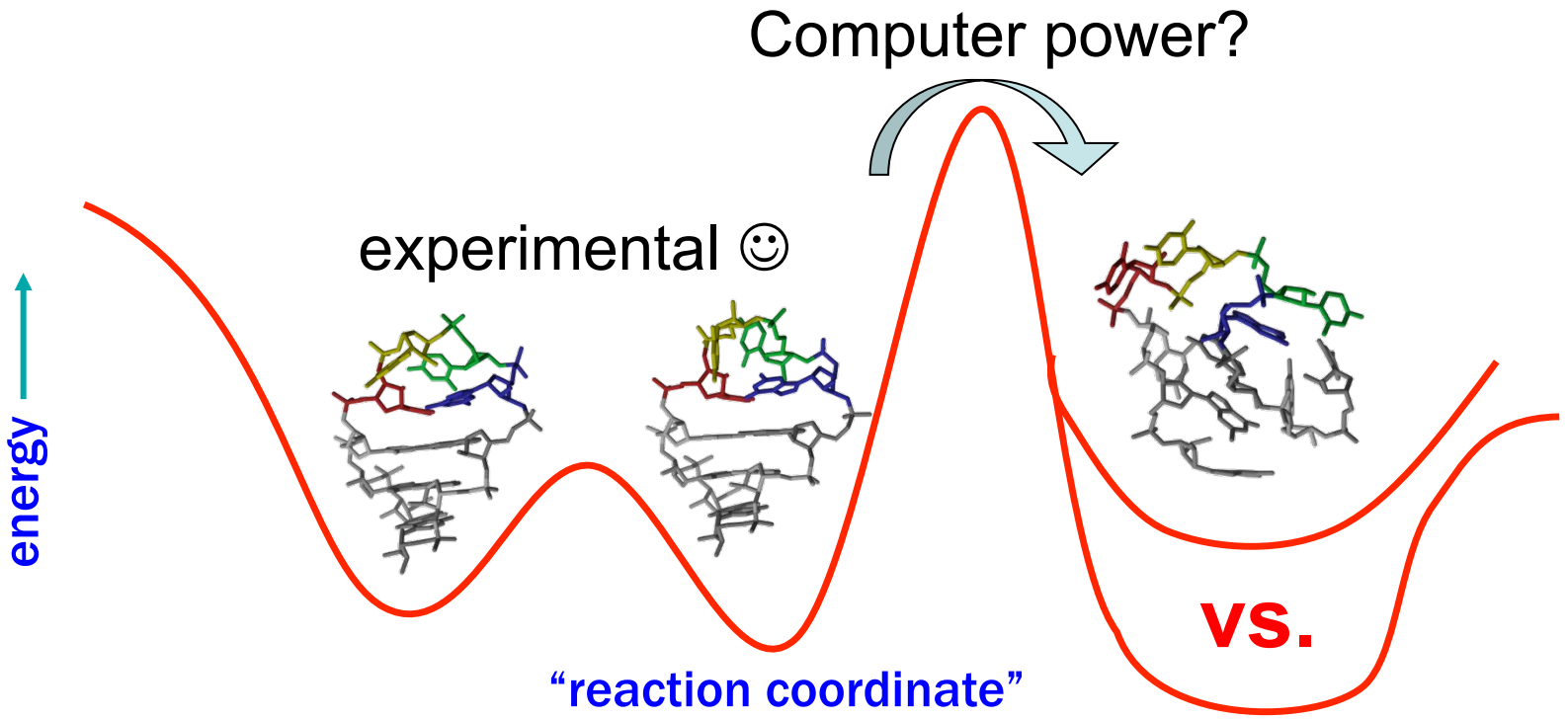
&

validation

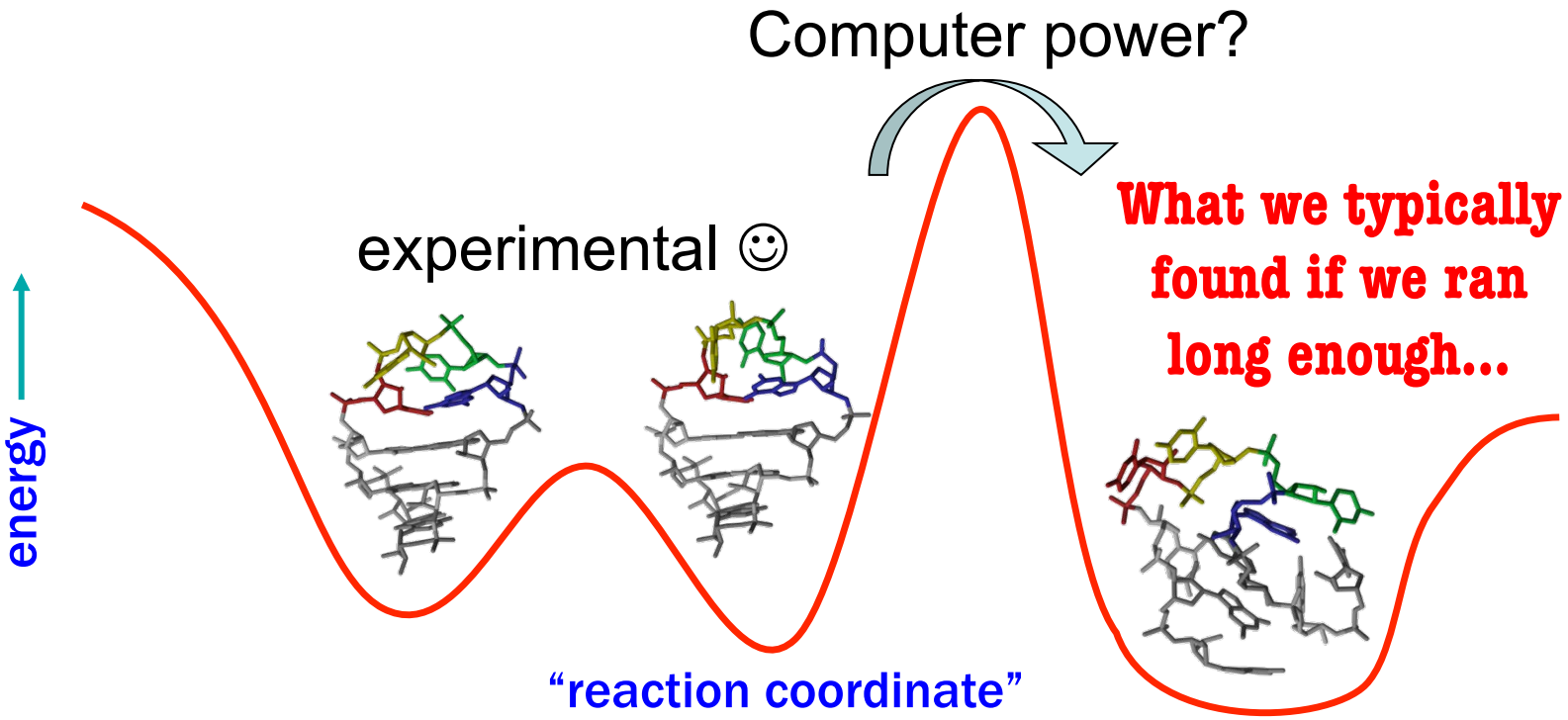
?

are the force fields reliable?
(free energetics, sampling, dynamics)

Short simulations stay near experimental structure; longer simulations invariably move away and often to unrealistic lower energy structures…



are the force fields reliable?
(free energetics, sampling, dynamics)



Updated nucleic acid force fields:

α,γ (γ=*trans*): parmbsc0

RNA ladder, X: X_{OL} (ff11, ff12) or $X_{Yildirim}$ or Chen/Garcia or bb/OPC

DNA ε/ζ: ε/ζ 2013 and/or X_{OL4} or parmbsc1 or CHARMM36

We can fully sample conformational ensembles!

(helices, tetranucleotides, RNA hairpins, …)

Simulating protein movements using Anton could aid drug design.

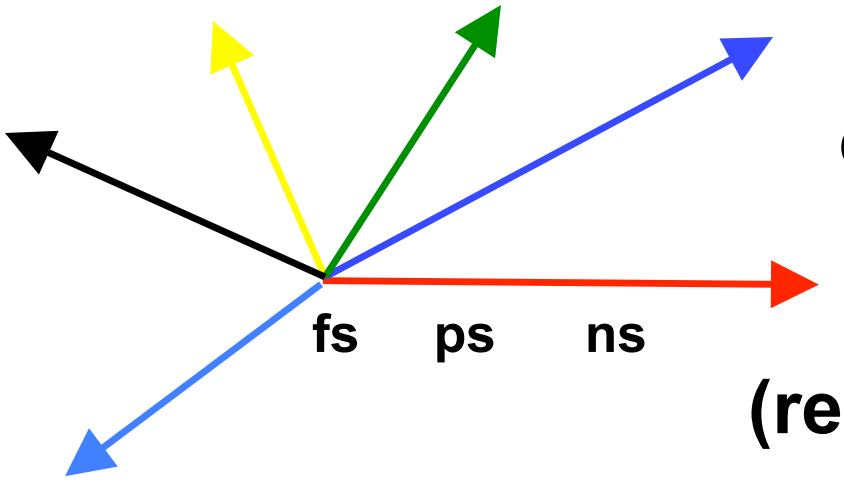
SCIENCE/AAAS

brute force – long contiguous in time MD
requires: special purpose / unique hardware

D.E. Shaw's Anton machine, 16 μs/day

or

AMBER on GPUs



ensembles of independent simulations (replica-exchange)

> 200 ns/day!

What did we learn?

- Ensembles of independent simulations show similar **convergence** properties with respect to the structure and dynamics of the internal part of a DNA helix

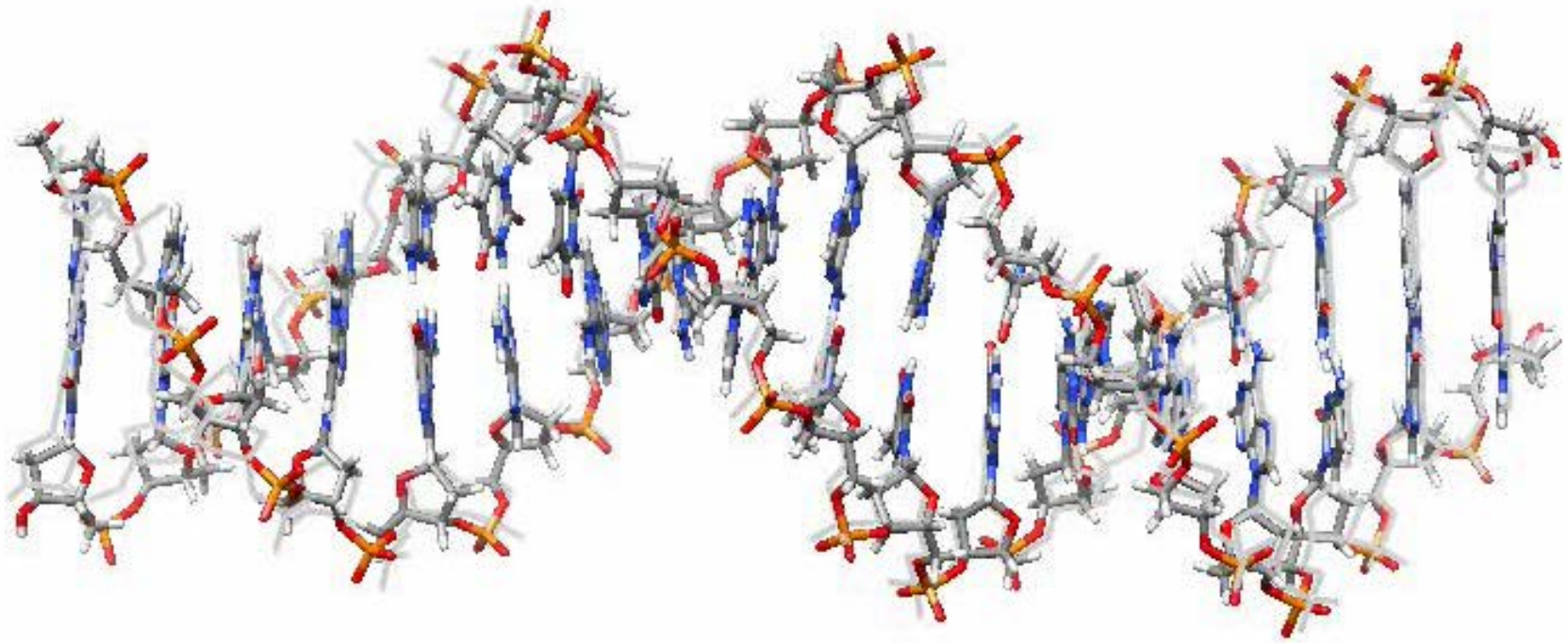
- Independent simulations (on special purpose hardware or GPUs or CPUs) give **reproducible** results

What are implications for DNA recognition?

- See: *Nature Comm*. 5:5152 (2014), *BBA* 1850, 1041-1058 (2015)

Convergence, force field and salt dependence in simulations of nucleic acids

d(GCACGAACGAACGAACGC) – Anton vs. GPUs

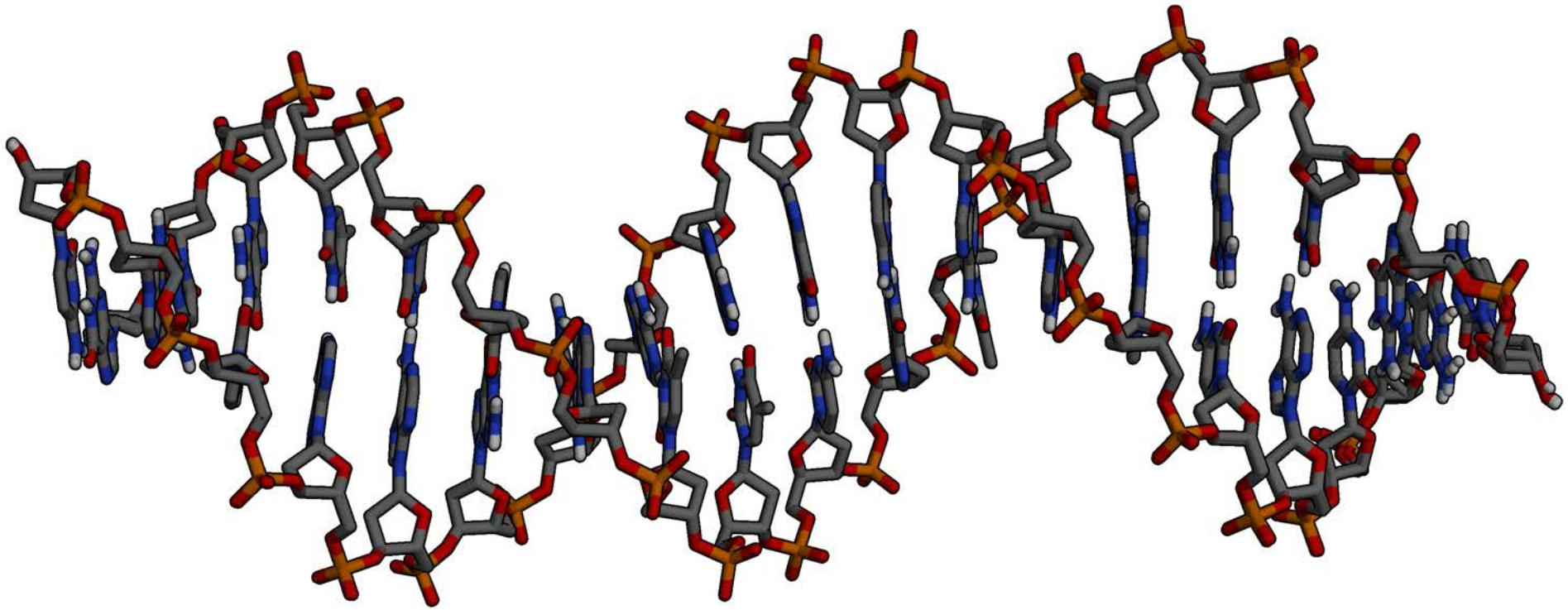


2 ns intervals (10 ns running average), render every 5th frame: ~10 us total time

~2010-2015

5 “average” structures overlayed @

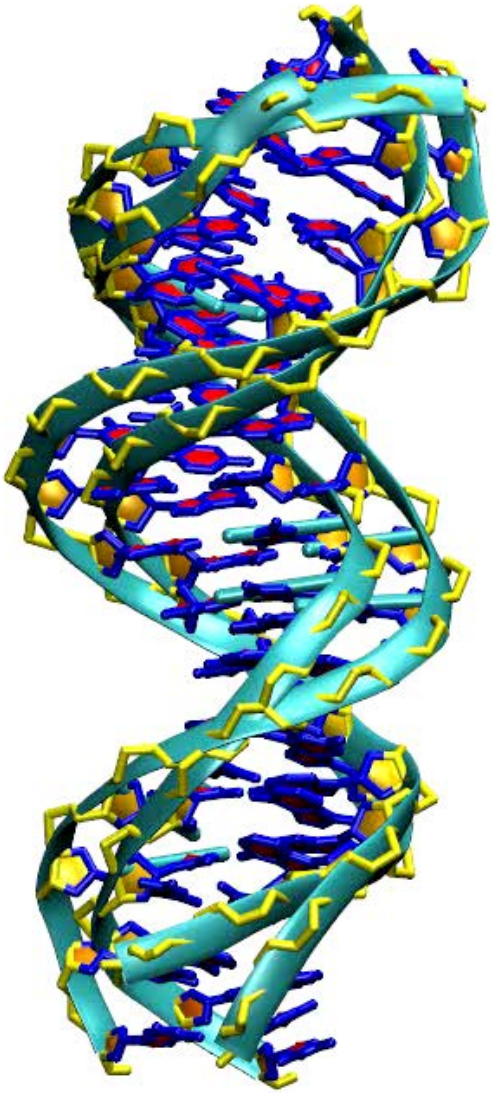
1.0-4.0 µs, 1.5-4.5 µs, 2.0-5.0 µs, 2.5-5.5 µs, 3.0-6.0 µs …
RMSd (0.028 Å) (0.049 Å) (0.076 Å) (0.160 Å)



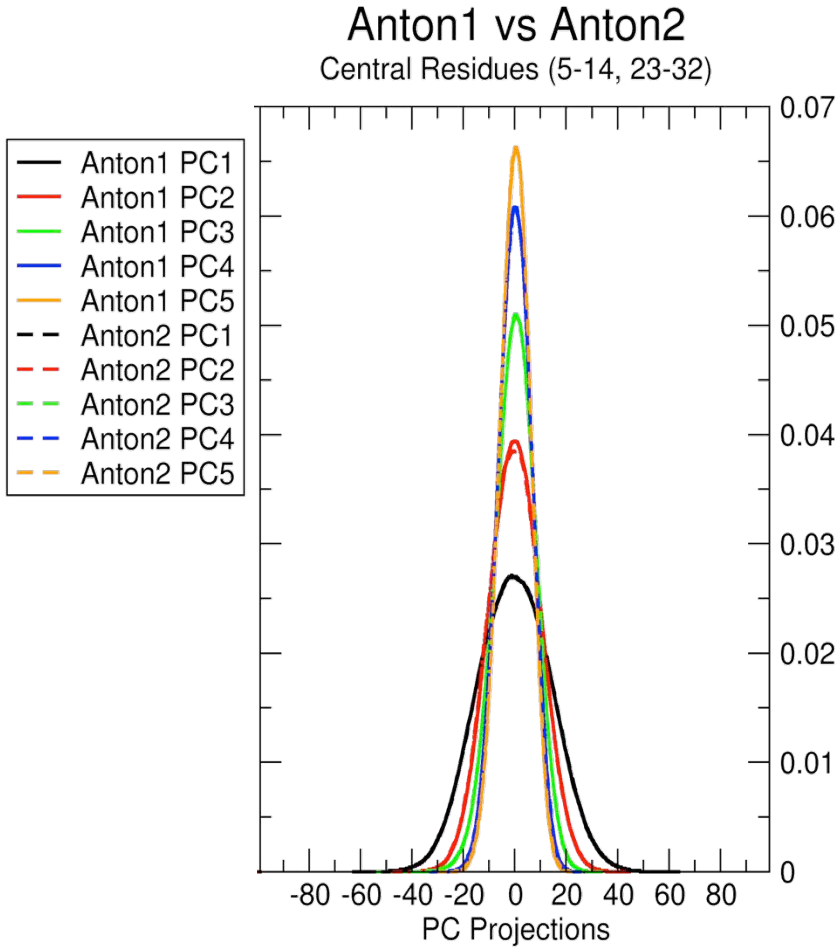
…this cannot be right, can it?
(breathing, bending, twisting, …)

Test for convergence within and between simulations: Dynamics

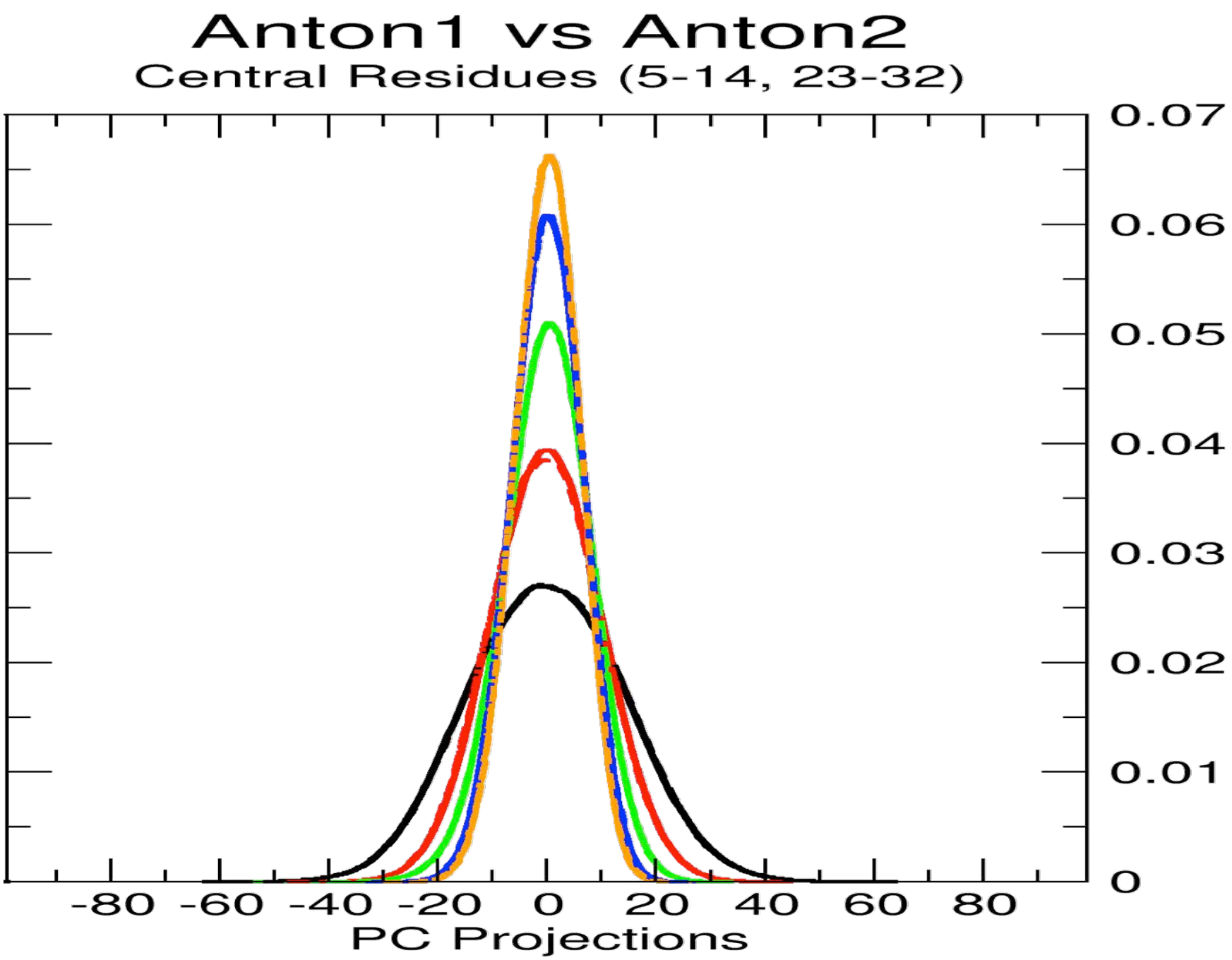
Principal components (or major modes of motion)



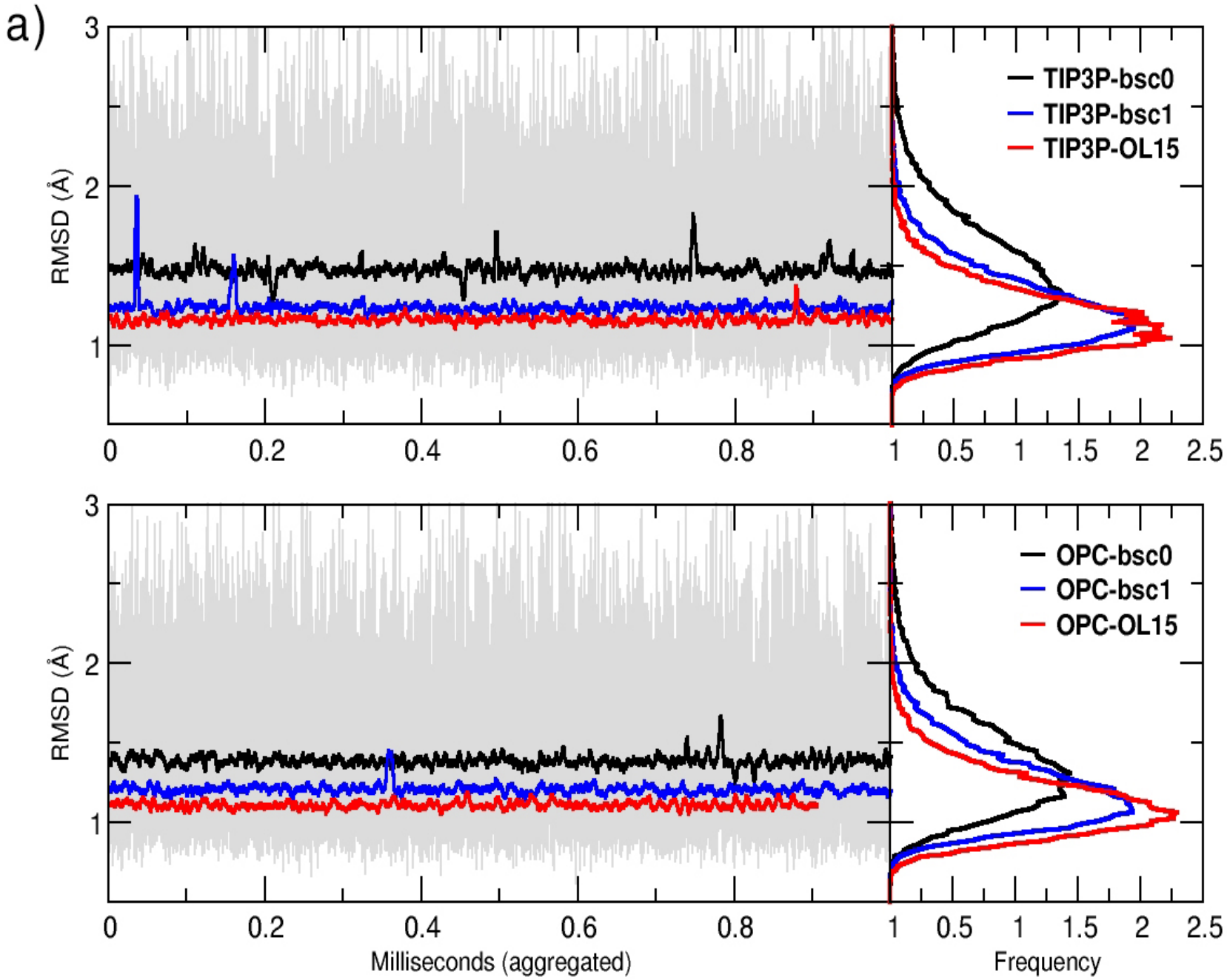
Visualization of the first two (dominant) modes of motion



Overlap of modes from independent simulations (internal helix)


Anton1 vs Anton2
Central Residues (5-14, 23-32)
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-80
-60
-40
-20
0
20
40
60
80
PC Projections

DDD, 100 independent MD trajectories, 11 μs, omit first 1 μs, aggregate

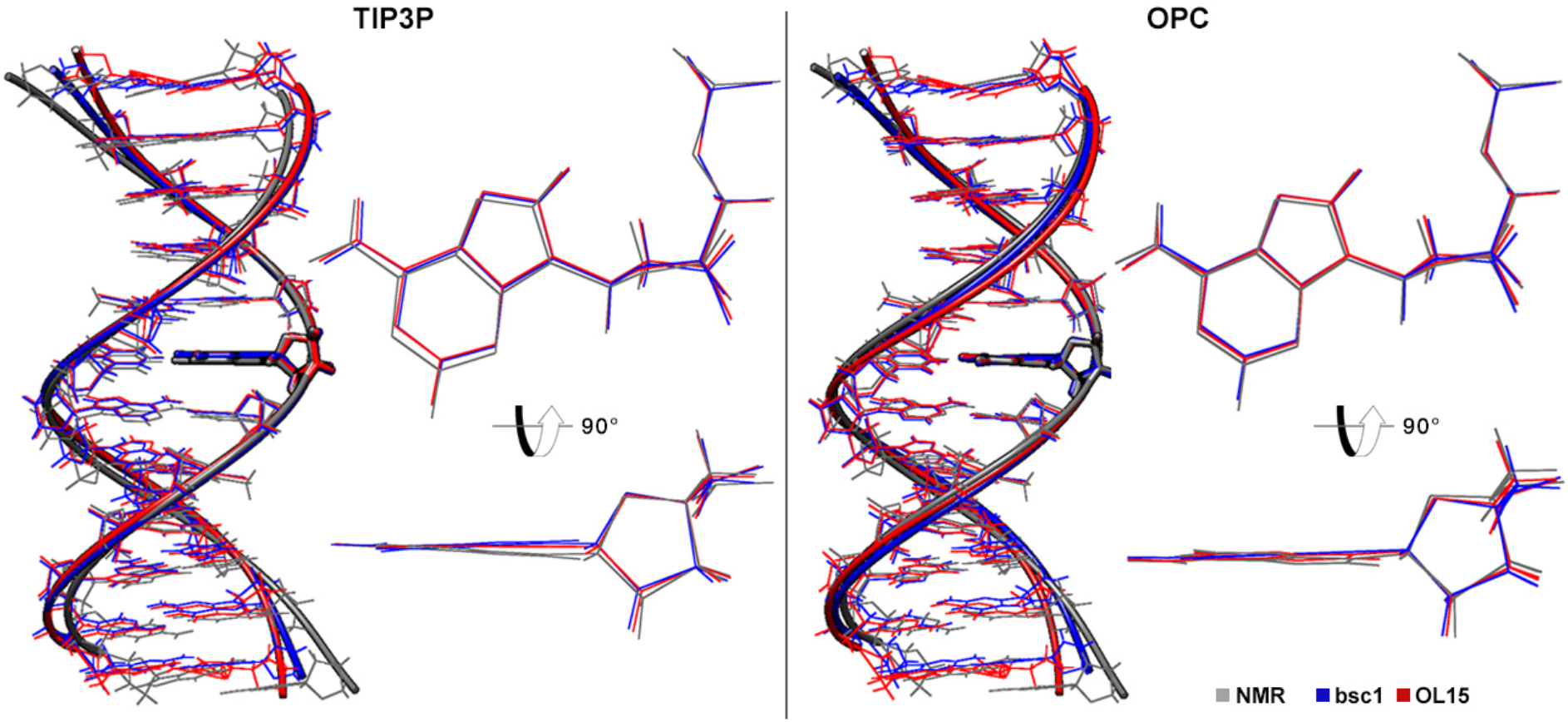


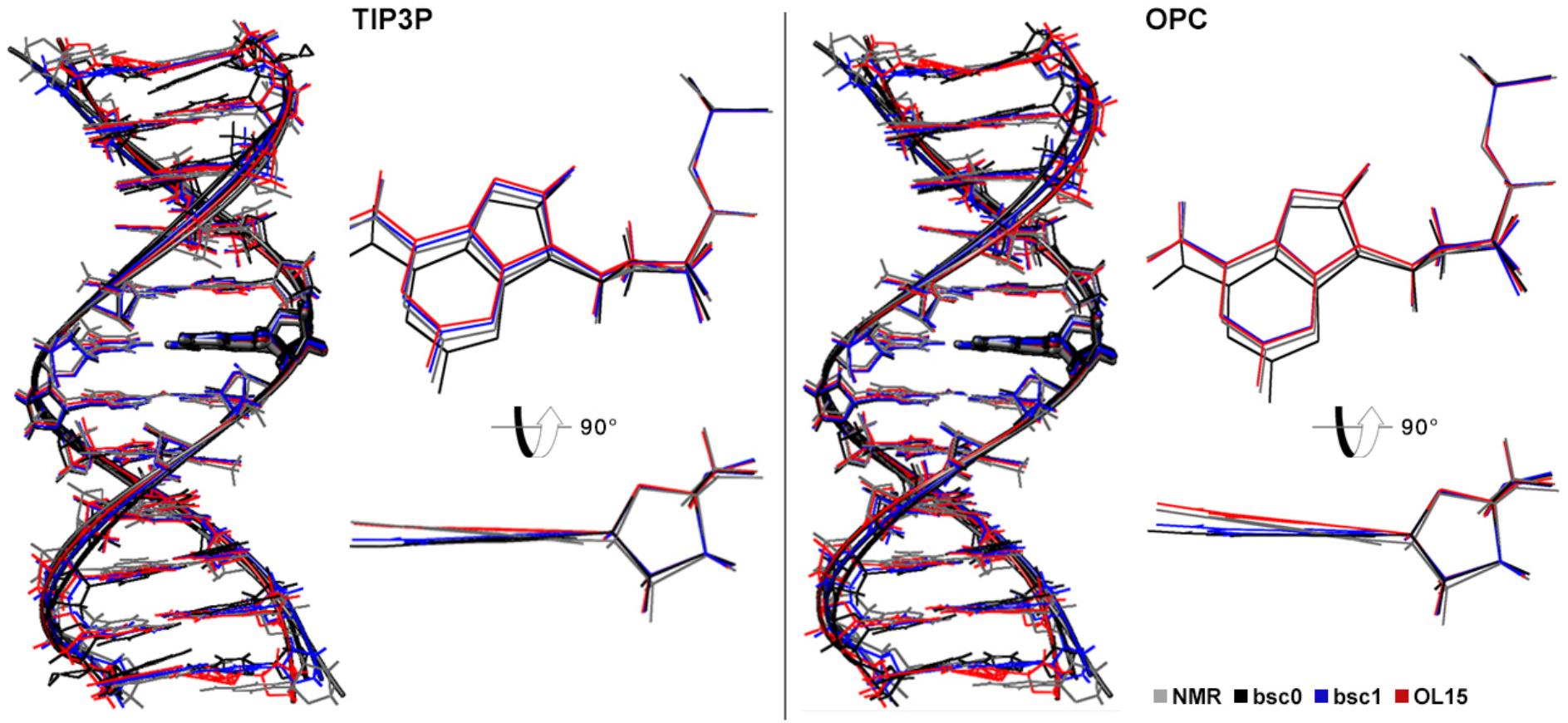
Root mean square (RMS) deviations (Å) of average structures from MD to NMR of the Dickerson dodecamer. The average structures from simulations were calculated over the full aggregated trajectories of each system (***100 independent MD trajectories, 11 μs, omit first 1 μs, aggregate – except C36 1.1 μs, omit first 200 ns***); the DDD NMR reference was an average of the models in the 1NAJ structure. RMS deviations were calculated over all heavy atoms of the internal eight base pairs.

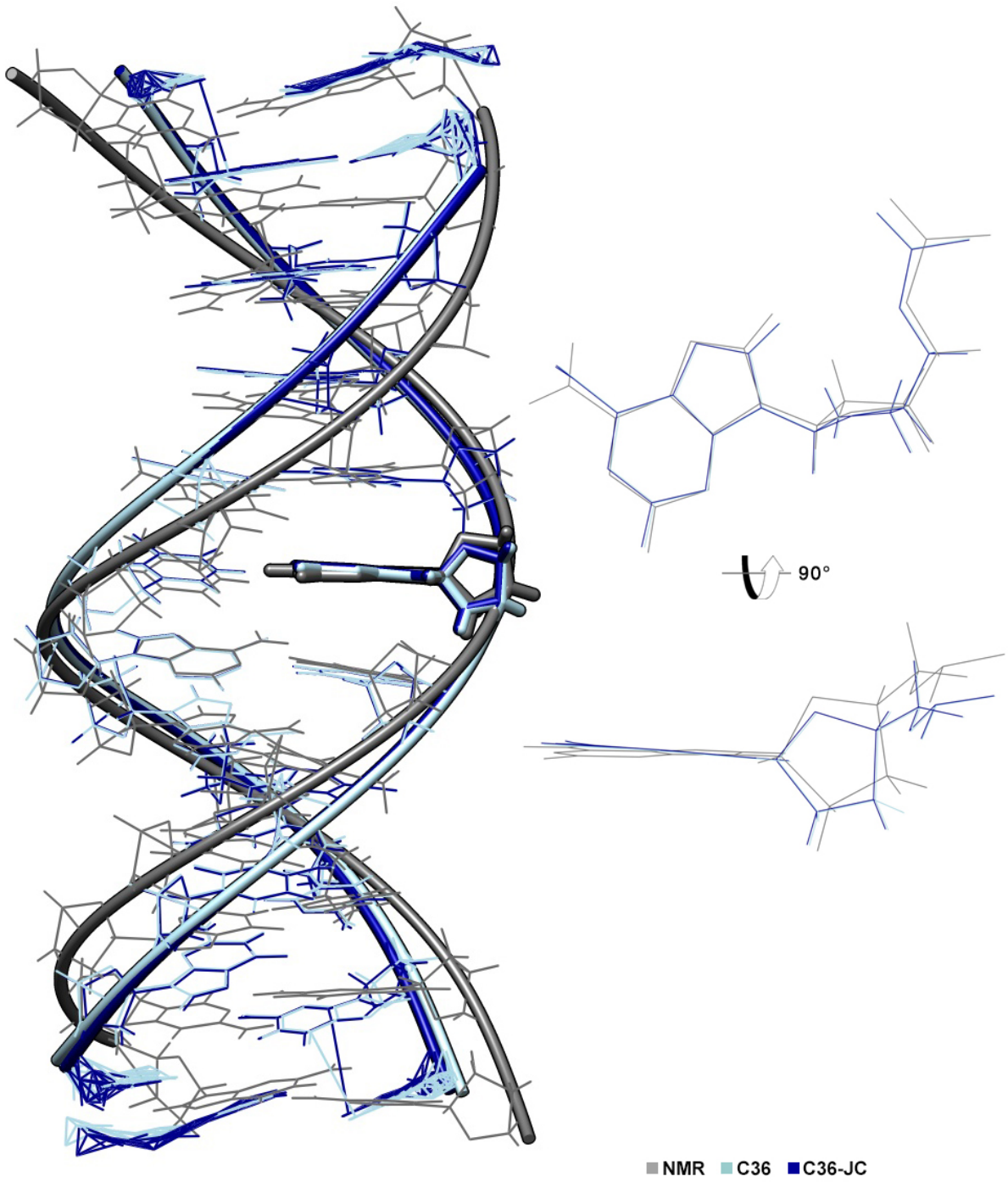
	bsc0	bsc1	OL15	CHARMM36	CHARMM36-JC
TIP3P	1.00	0.64	0.53	1.29	1.30
OPC	0.91	0.61	0.44		

Wow!

Deviation to experiment…

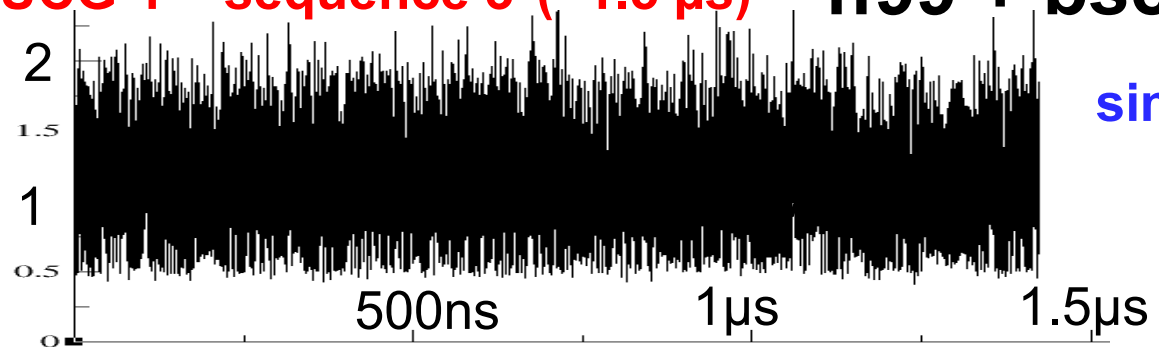

TIP3P
OPC
90°
90°
NMR
bsc1
OL15


TIP3P
90°
OPC
90°
NMR
bsc0
bsc1
OL15

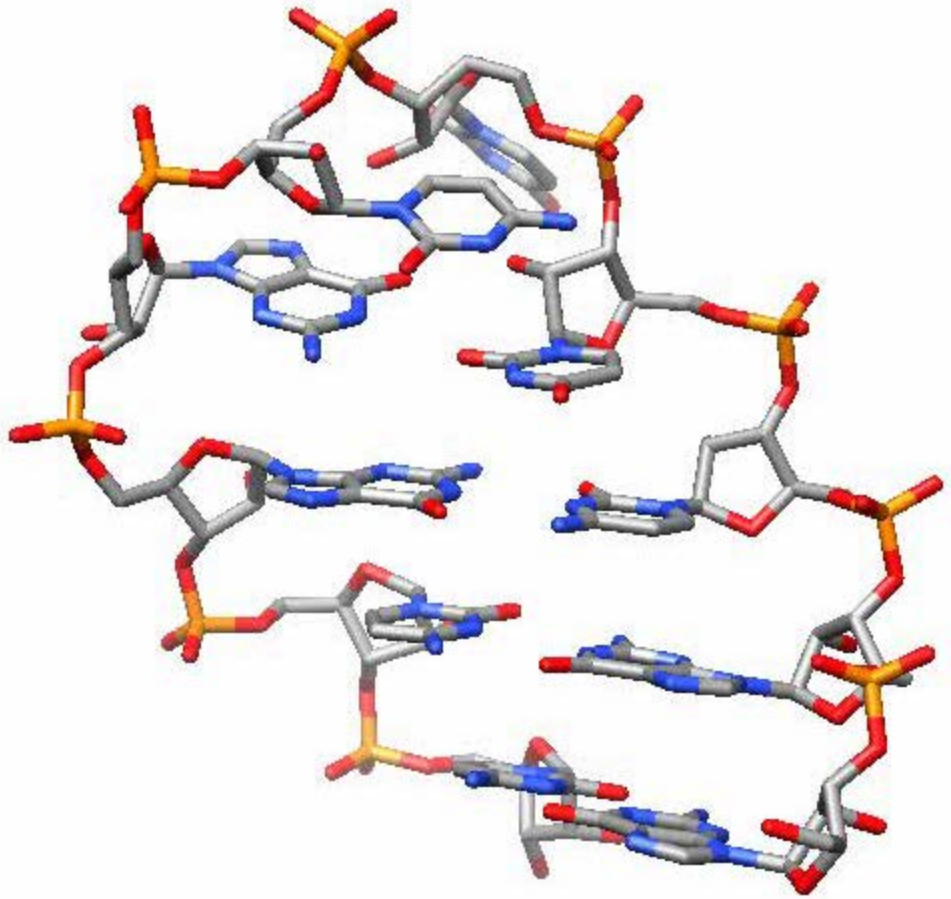

90°
NMR
C36
C36-JC

UUCG-1 – sequence 3 (~1.5 μs)

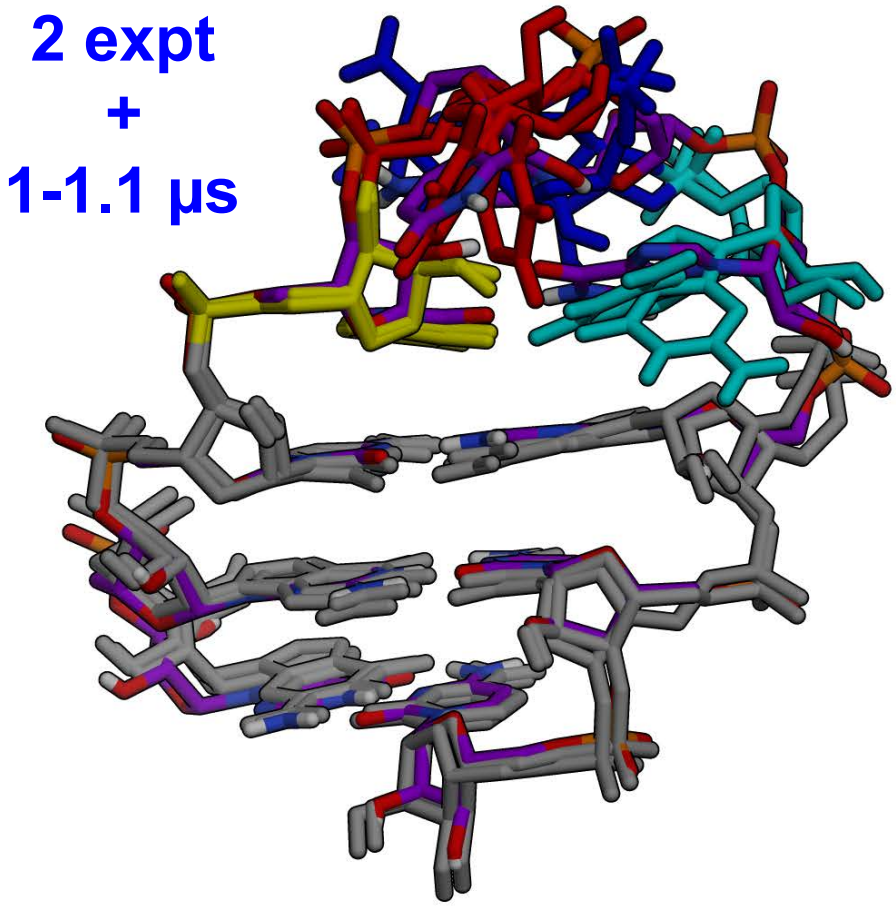
ff99 + bsc0 + OL χ fix

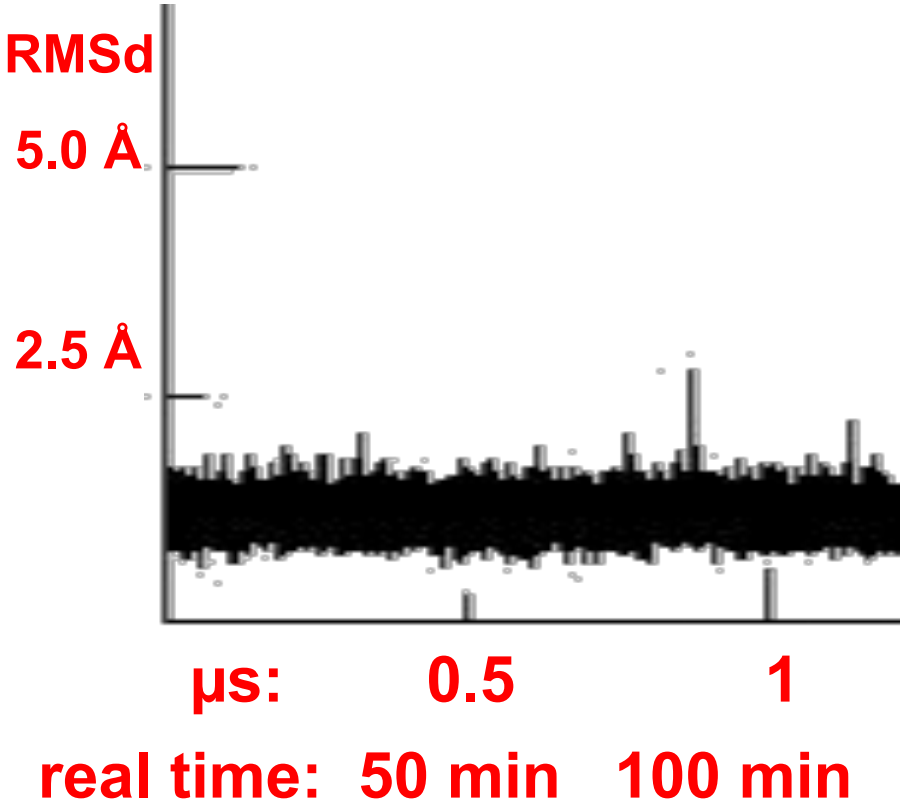


simulated w/out restraints, modern force field, explicit solvent



RNA UUCG tetraloop (ff99bsc0 + OL X) on Anton @ PSC:



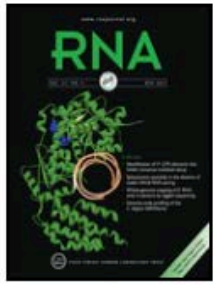


Initial tests: RNA tetraloop

Structural fidelity and NMR relaxation analysis in a prototype RNA hairpin.

Giambaşu GM, York DM, **Case** DA.

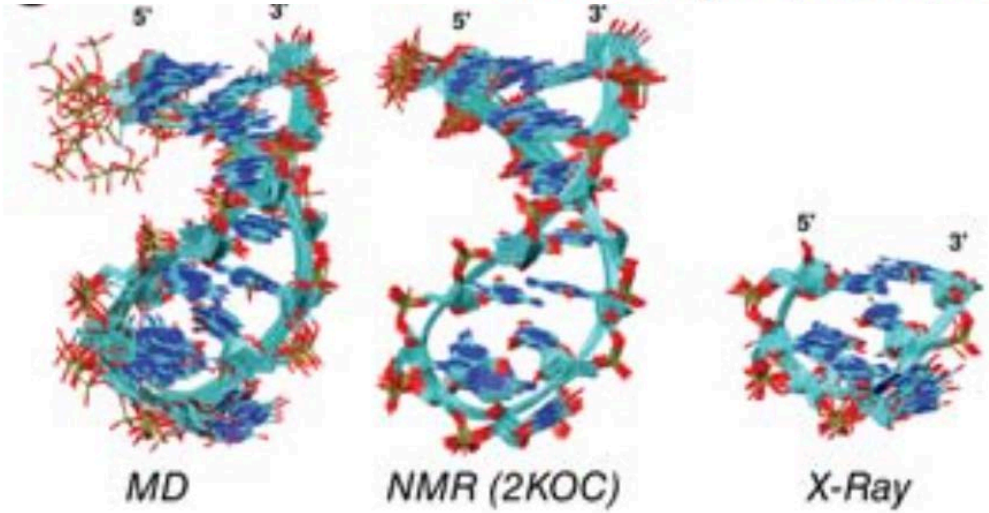
RNA. 2015 May;21(5):963-74. doi: 10.1261/rna.047357.114. Epub 2015 Mar 24.
PMID: 25805858

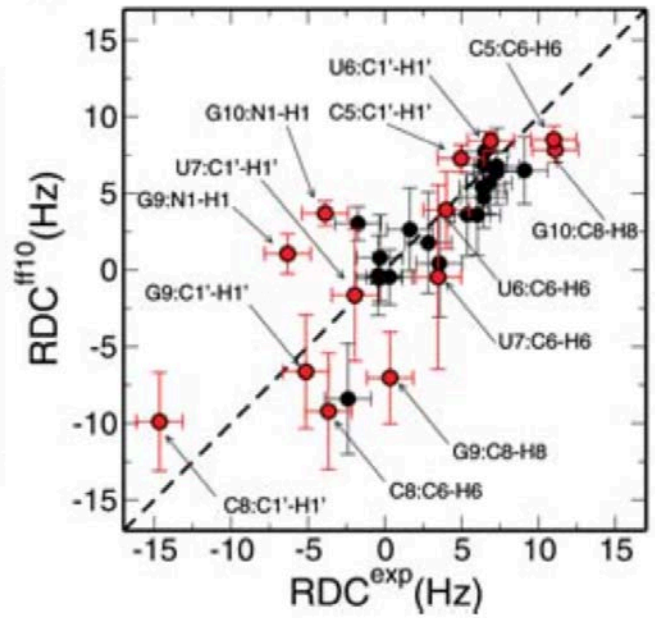


Full Text

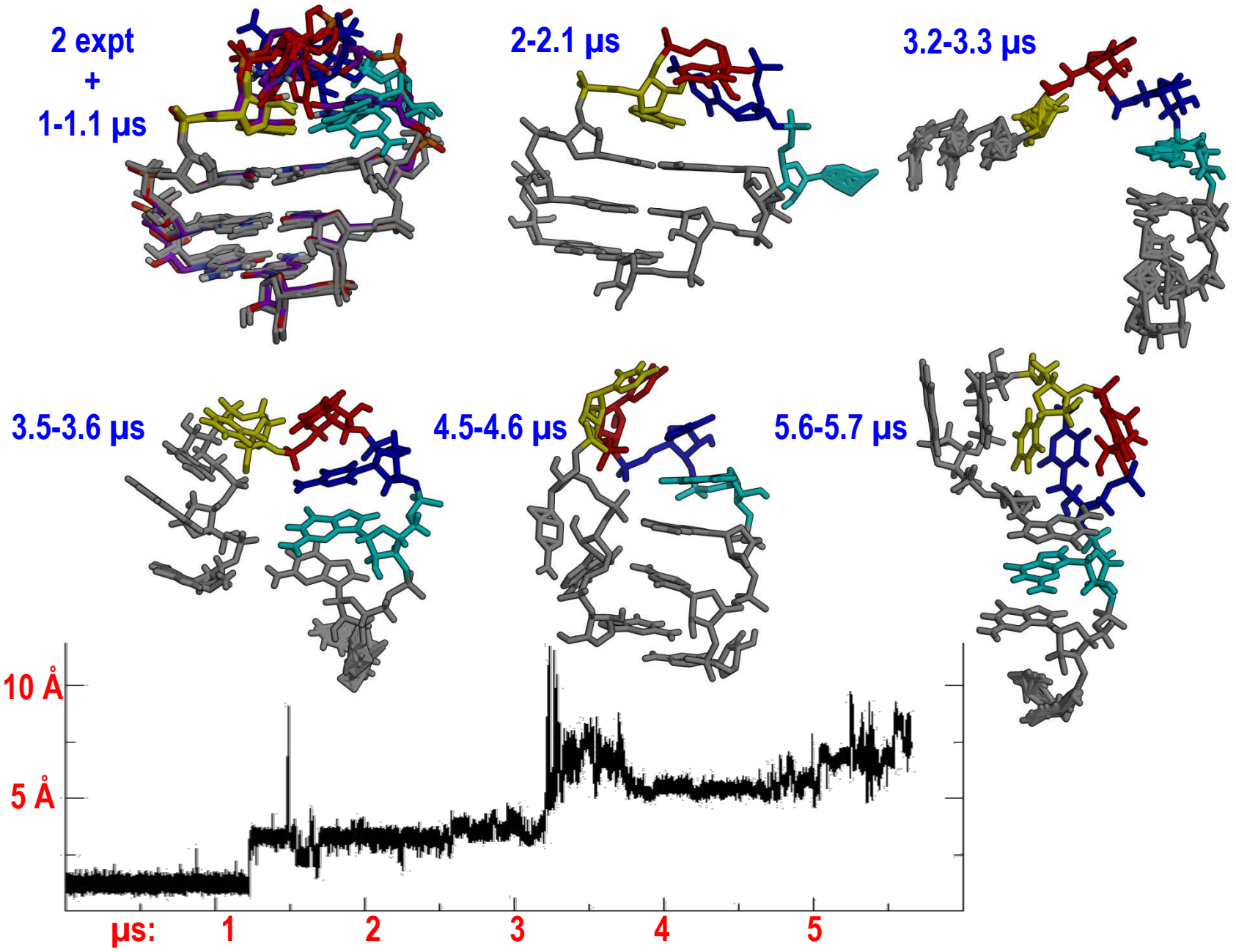
Structural fidelity and NMR relaxation analysis in a prototype RNA hairpin

RNA May 2015 21: 963-974; Published

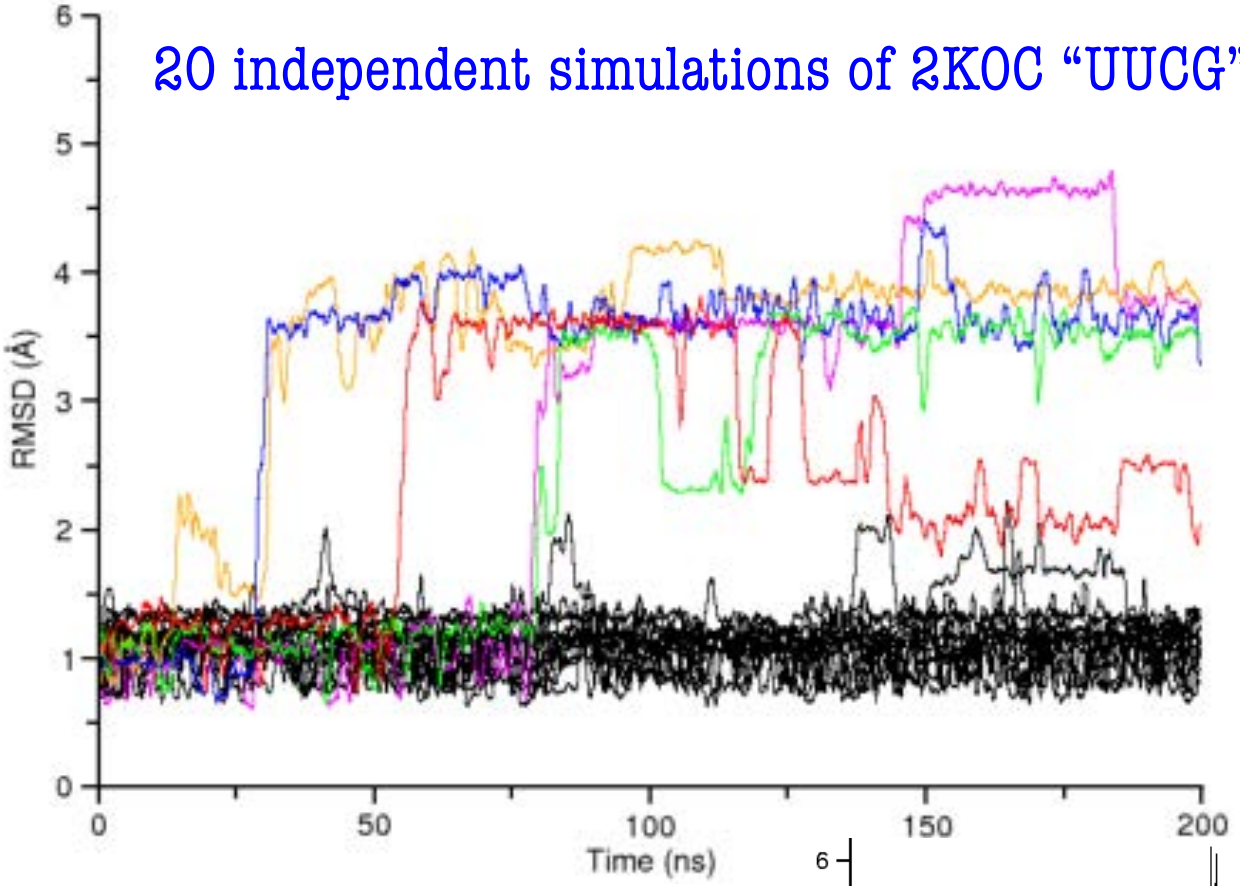




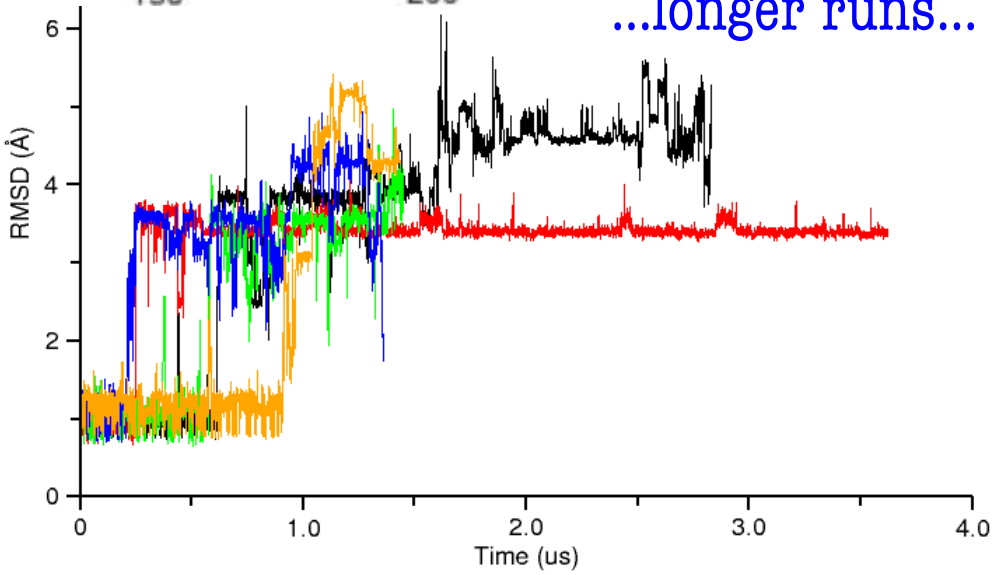
RNA UUCG tetraloop (ff99bsc0 + OL X):



20 independent simulations of 2KOC “UUCG” tetraloop



...longer runs...



…before we get to some “issues”, what can we say about ion dependent RNA structure?

vsrSL5 & GAC

J. Phys. Chem. B (2015)

Stem-Loop V of Varkud Satellite RNA Exhibits Characteristics of the Mg^{2+} Bound Structure in the Presence of Monovalent Ions

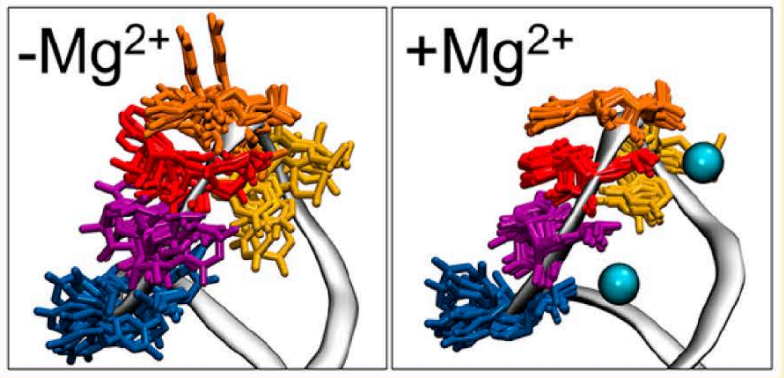
Christina Bergonzo,[†] Kathleen B. Hall,[‡] and Thomas E. Cheatham, III*[,†]

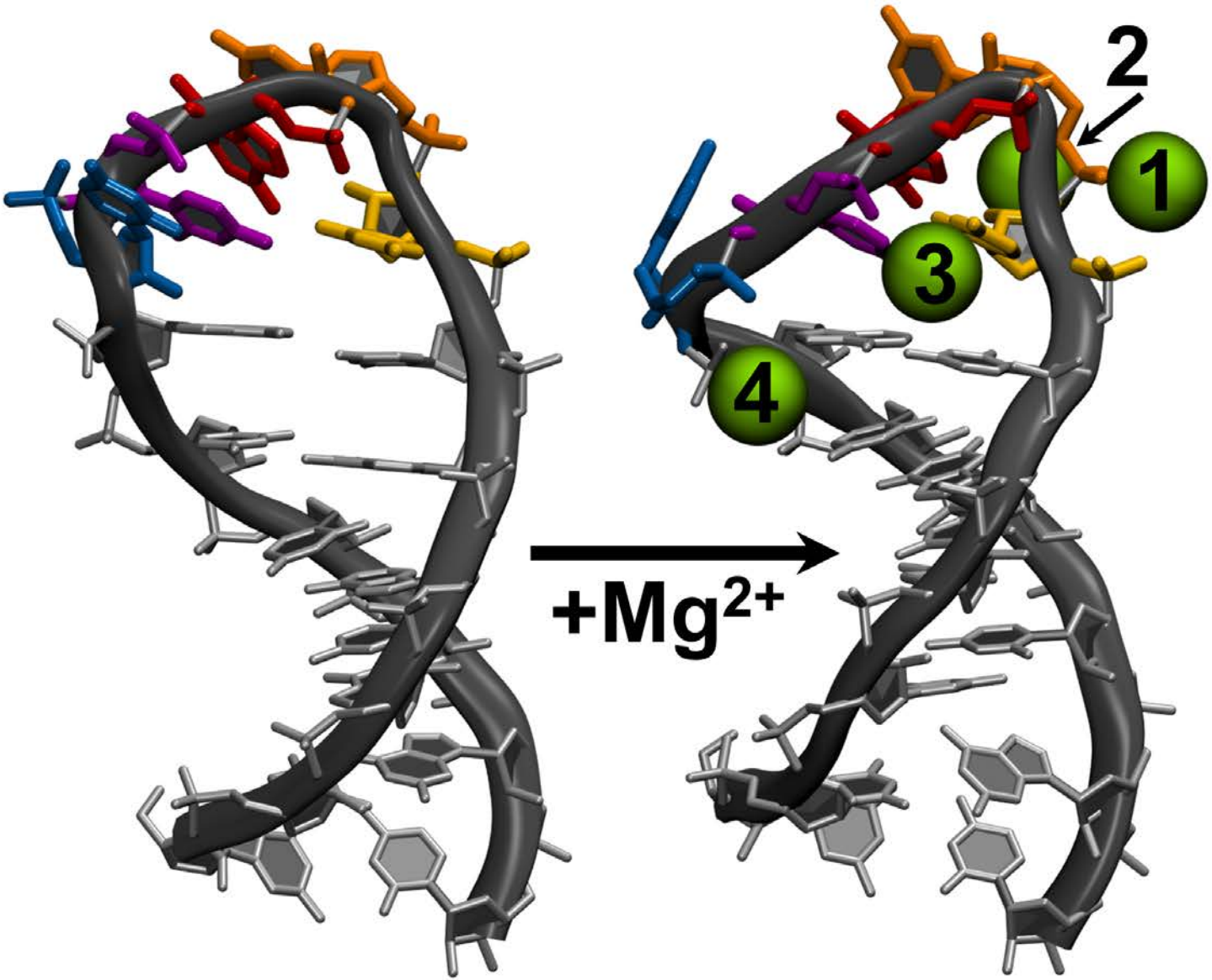
[†]Department of Medicinal Chemistry, College of Pharmacy, University of Utah, Salt Lake City, Utah 84112, United States

[‡]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110, United States

Supporting Information

ABSTRACT: The Varkud Satellite RNA contains a self-cleaving ribozyme that has been shown to function independently of its surroundings. This 160 nucleotide ribozyme adopts a catalytically active tertiary structure that includes a kissing hairpin complex formed by stem-loop I and stem-loop V (SLV). The five-nucleotide 5′-rUGACU loop of the isolated SLV has been shown to adopt a Mg^{2+}-dependent U-turn structure by solution NMR. This U-turn hairpin is examined here by molecular dynamics simulations in the presence of monovalent and divalent ions. Simulations confirm on an all-atom level the hypotheses for the role of the Mg^{2+} ions in stabilizing the loop, as well as the role of the solvent exposed U_{700} base. Additionally, these simulations suggest the Mg^{2+}-free stem-loop adopts a wide range of structures, including energetically favorable structures similar to the Mg^{2+}-bound loop structure. We propose this structure is a "gatekeeper" or precursor to Mg^{2+} binding when those ions are present.

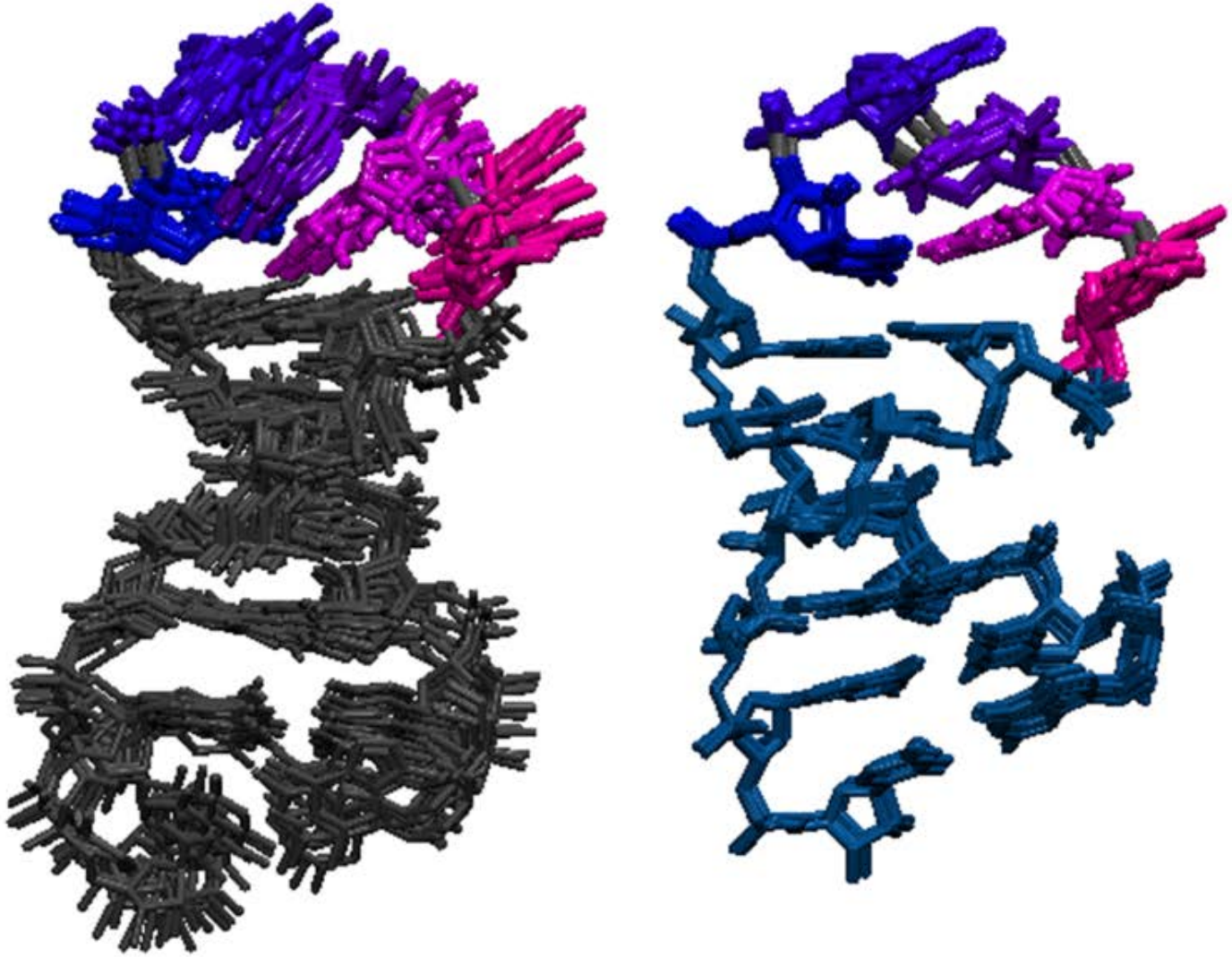




5’-GCGAG-$U_6U_7G_8A_9C_{10}U_{11}A_{12}$-CUCGC-3’

decoy: 1TBK ↔ 1YN2 ± Mg^{2+}

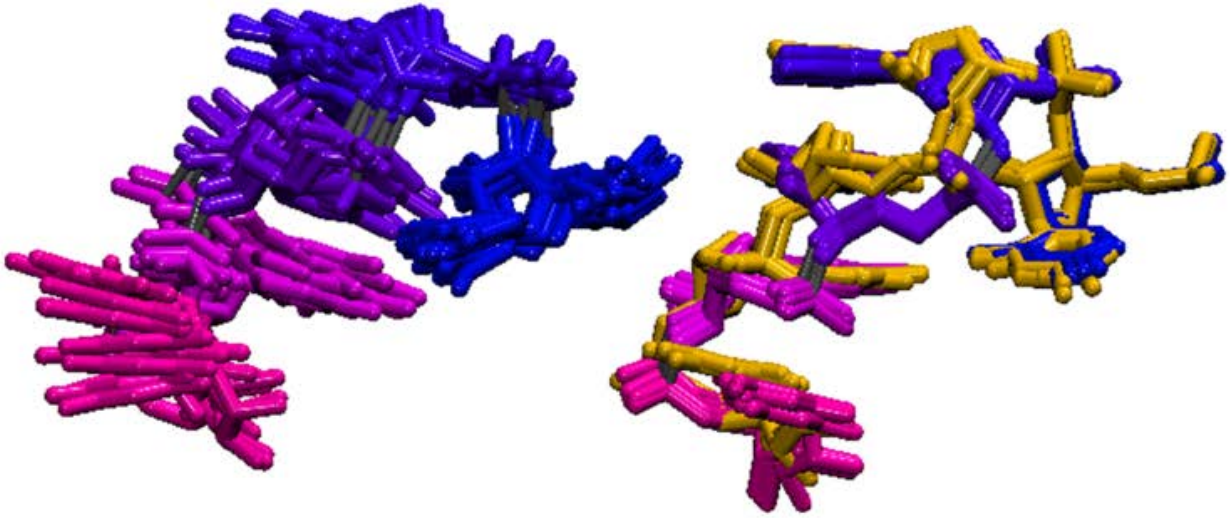
-Mg^{2+} deviates from NMR structure: re-refine…

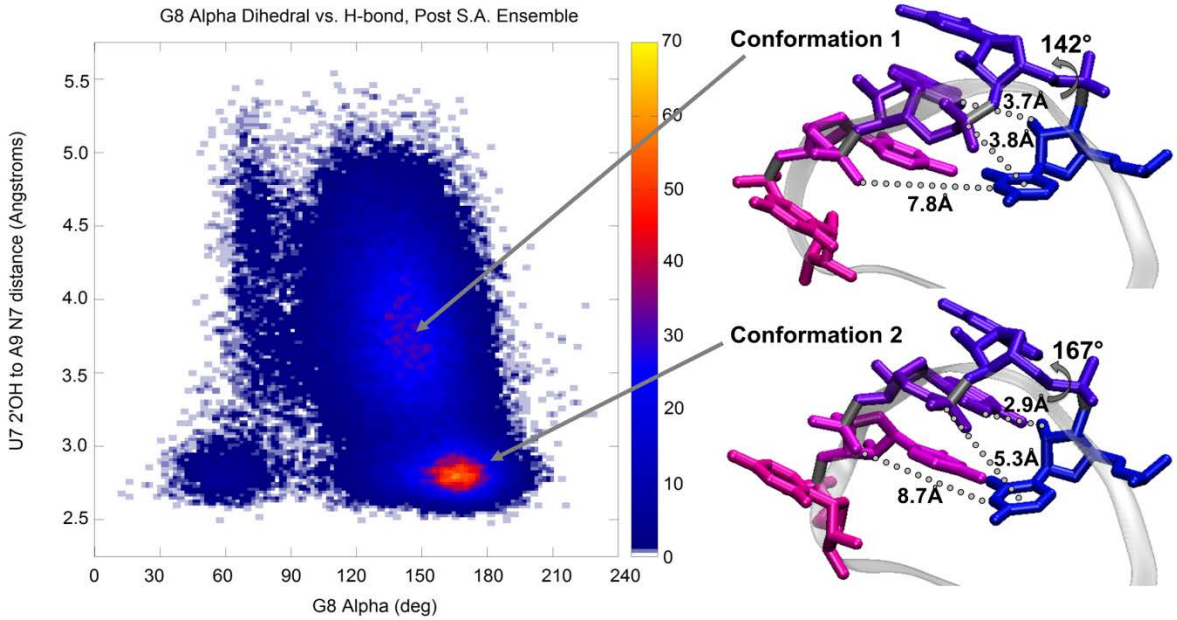


original NMR **re-refined NMR**

decoy: 1TBK ↔ 1YN2 ± Mg^{2+}

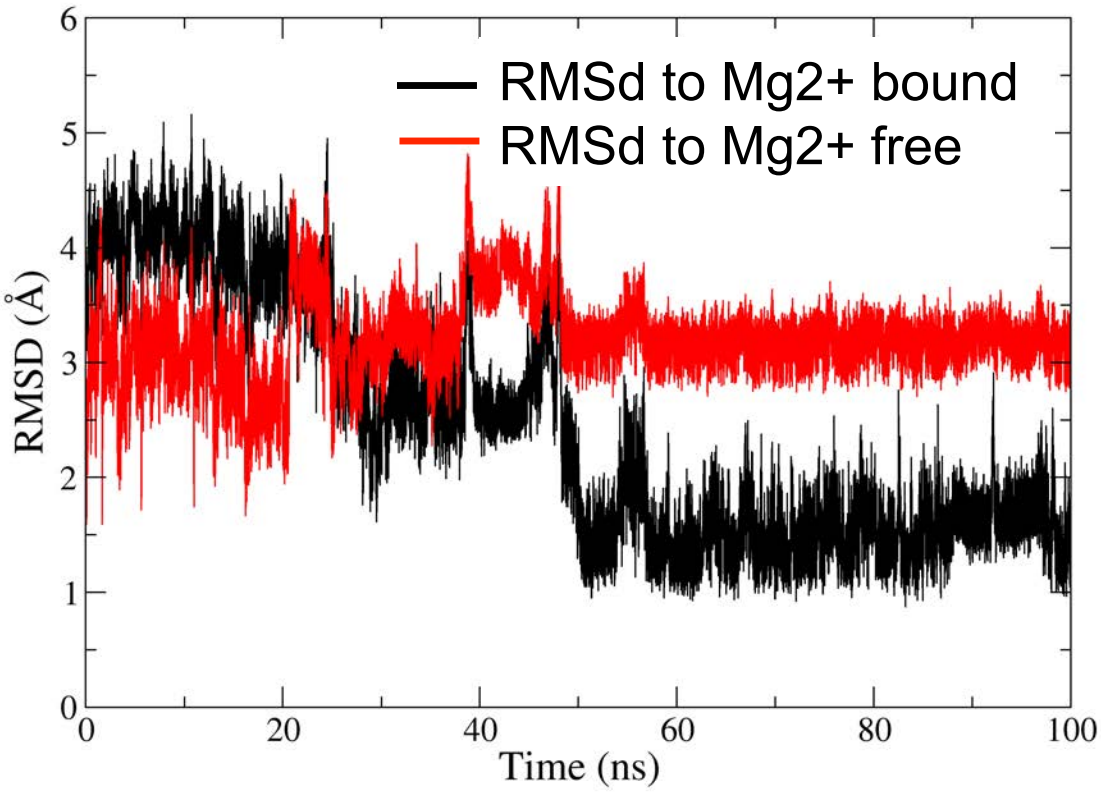
-Mg^{2+} deviates from NMR structure: re-refine…





We're seeing some progress!!!

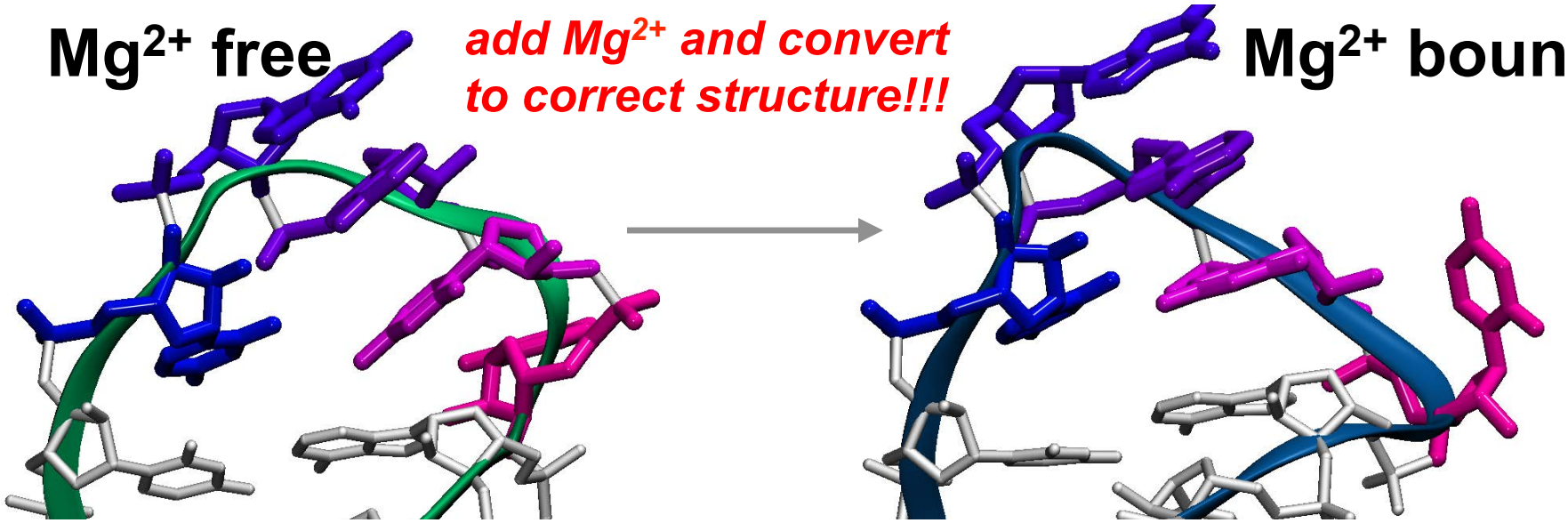
(vsrSL5)



Mg^{2+} free

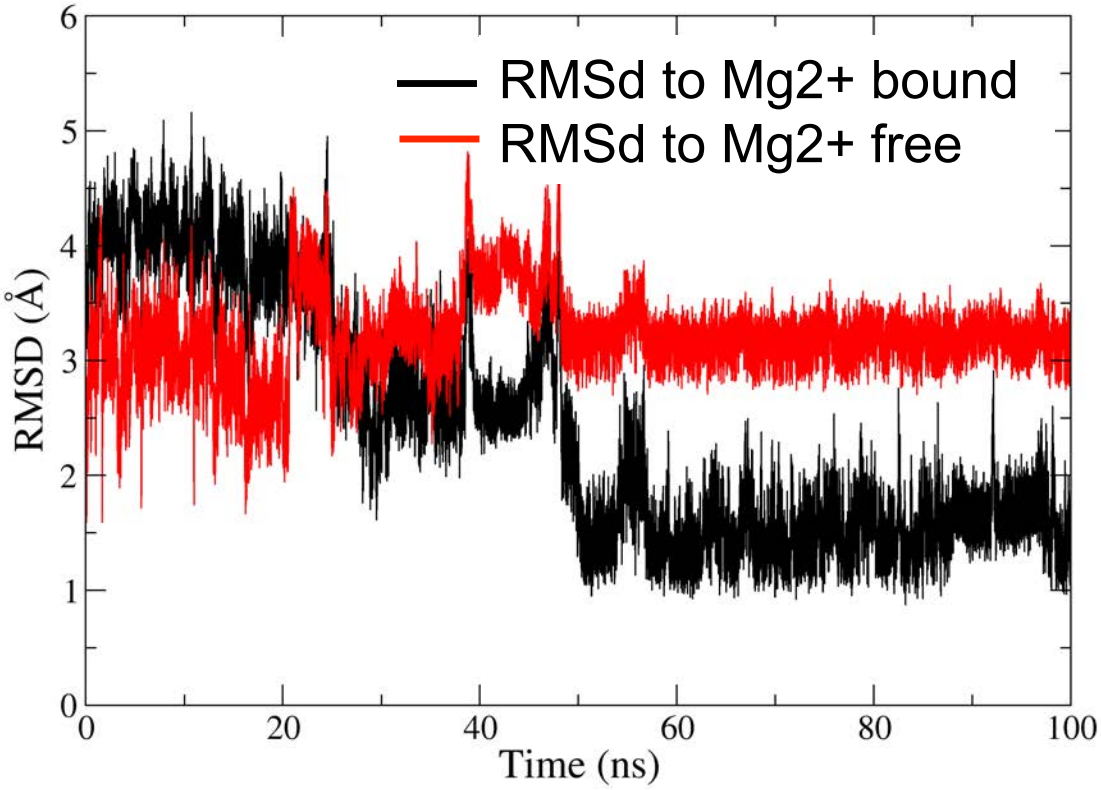
add Mg^{2+} and convert to correct structure!!!

Mg^{2+} bound



We're seeing some progress!!!

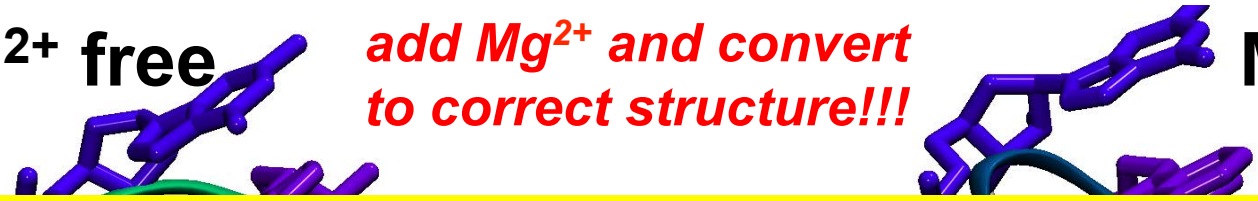
(vsrSL5)

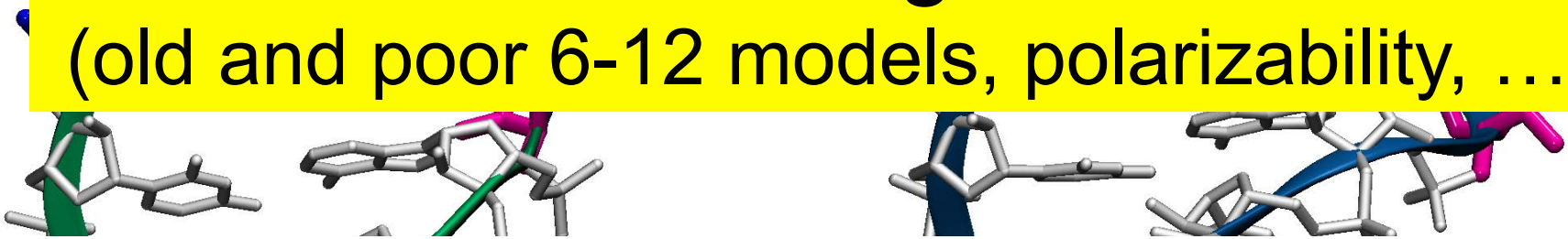


Mg^{2+} free

add Mg^{2+} and convert to correct structure!!!

Mg^{2+} bound



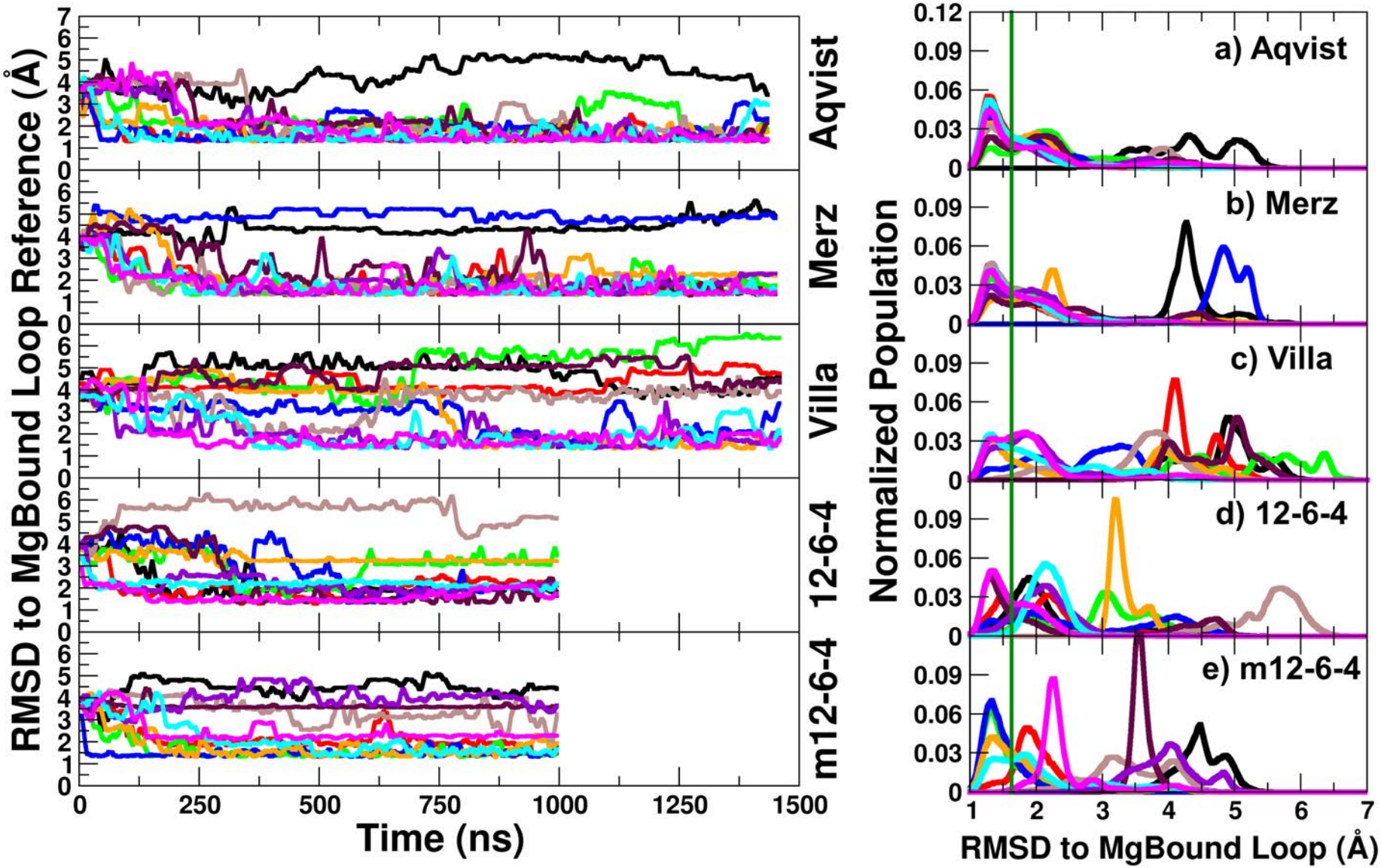


What about the Mg^{2+} ion model?

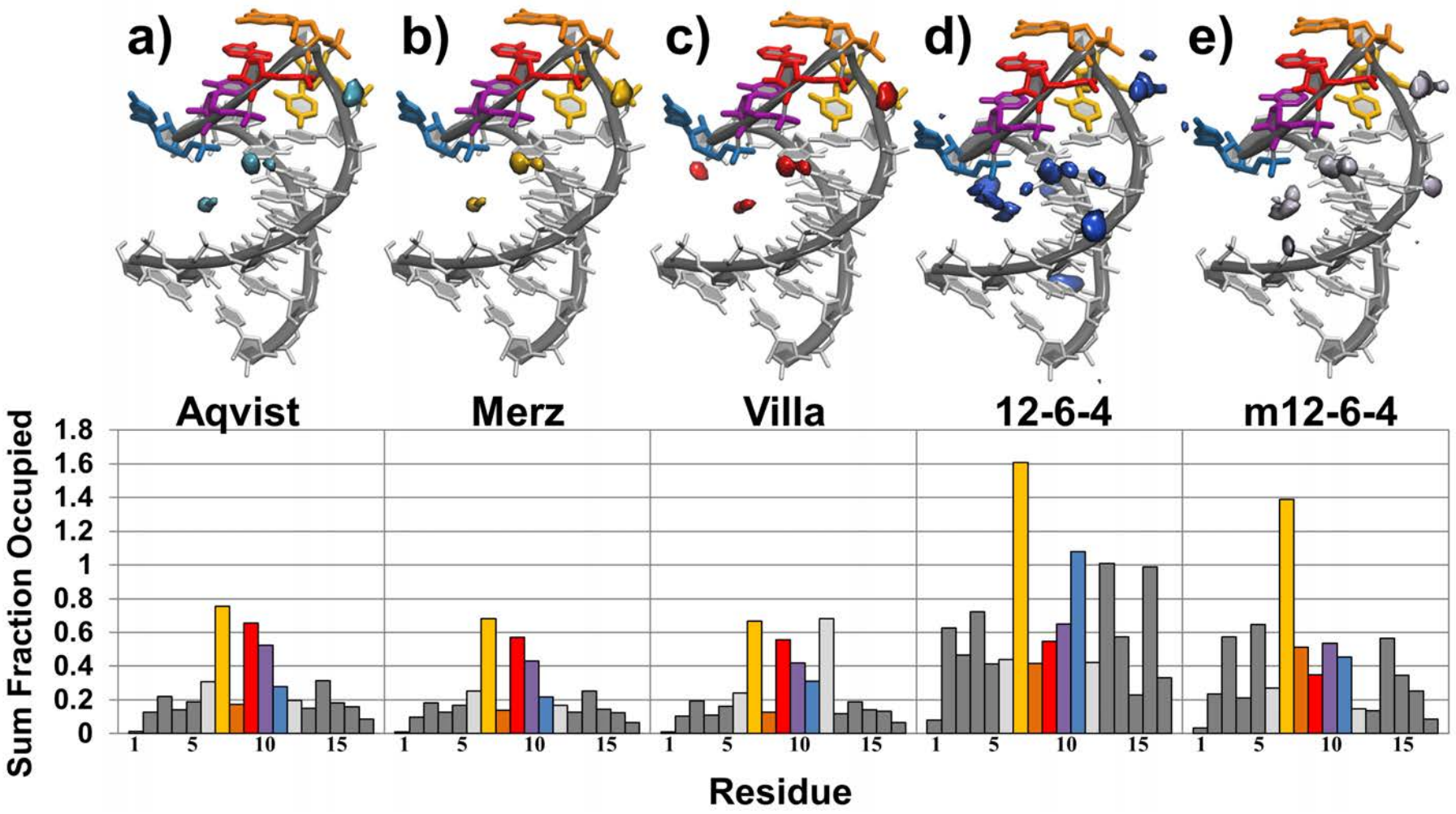
(old and poor 6-12 models, polarizability, …)

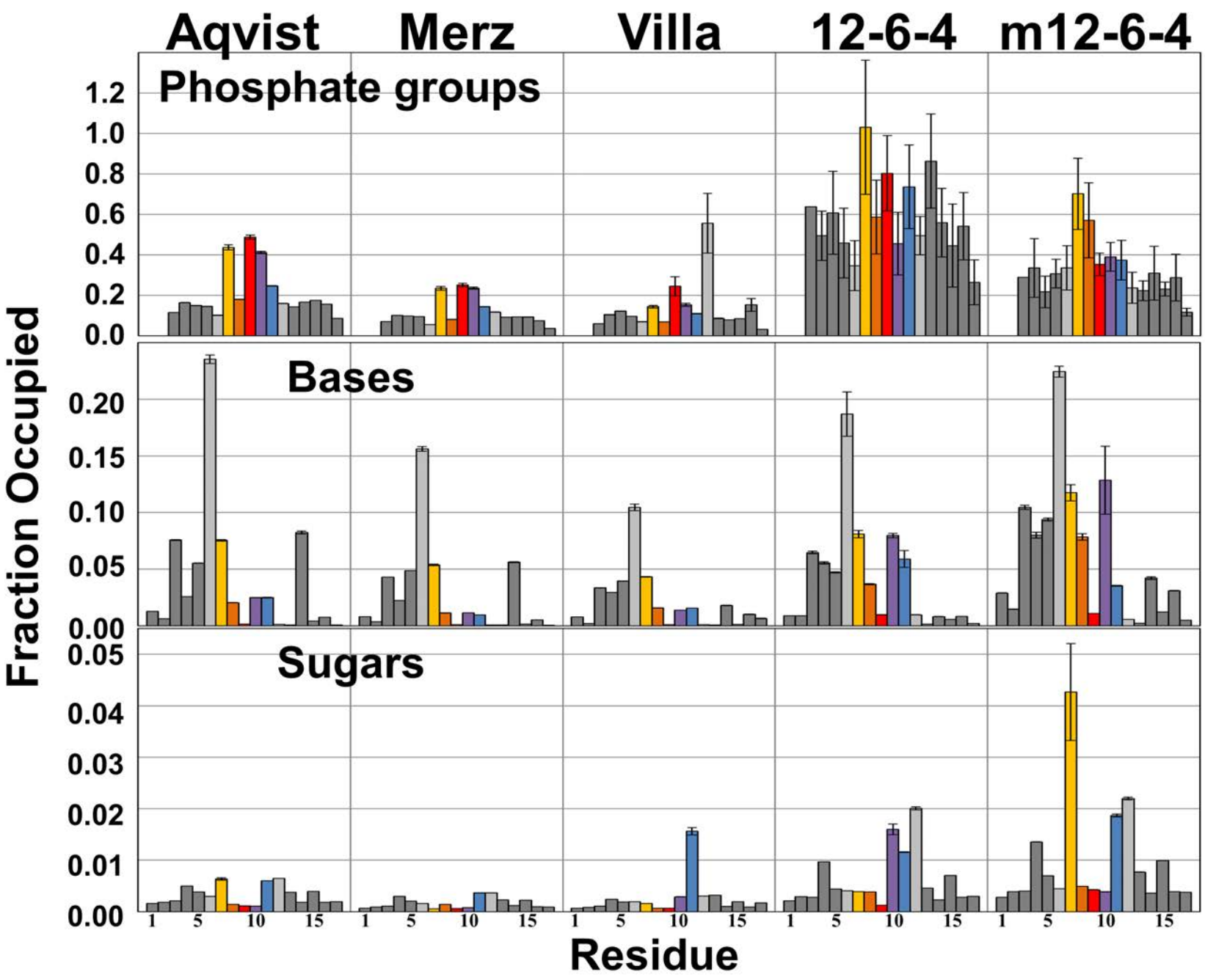
	$R_{min}/2$ (Å)	ε (kcal/mol)	C12	C6	C4
Aqvist	0.7926	0.8947	225	28.4	n/a
Merz	1.360	0.01020237	1673	8.26	n/a
Villa	1.5545	0.00295	2400	5.32	n/a
12-6-4/m12-6-4	1.437	0.02258	7171	25.4	4.42

lower values mean closer to Mg^{2+} bound structure

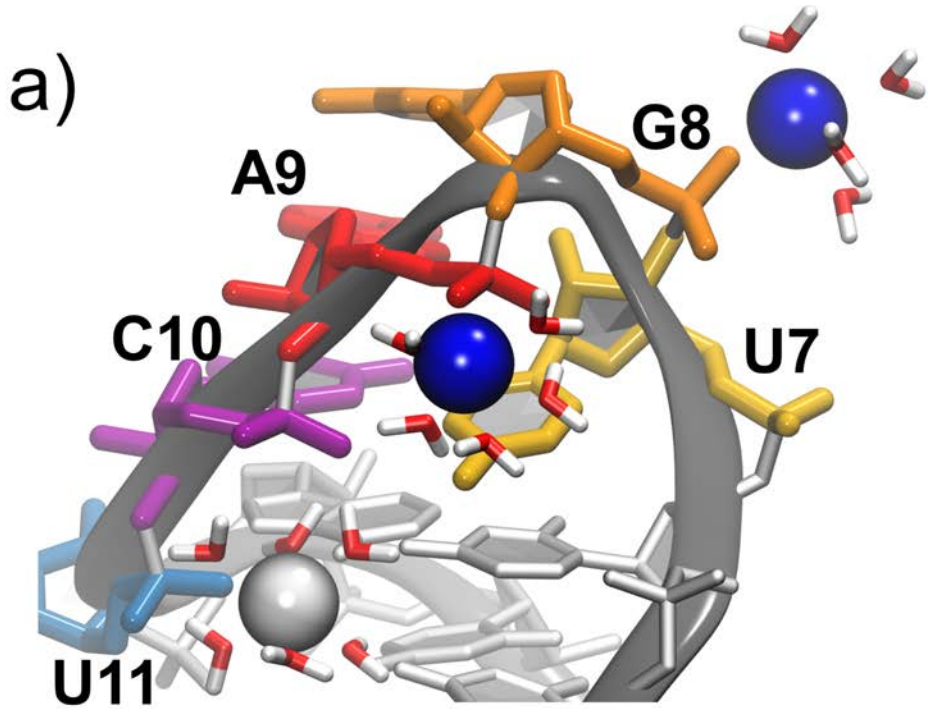


why is MD so short?


a)
b)
c)
d)
e)
Aqvist
Merz
Villa
12-6-4
m12-6-4
Sum Fraction Occupied
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
5
10
15
Residue


Aqvist
Merz
Villa
12-6-4
m12-6-4
Phosphate groups
Bases
Sugars
Fraction Occupied
Residue
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.20
0.15
0.10
0.05
0.00
0.05
0.04
0.03
0.02
0.01
0.00
1
5
10
15

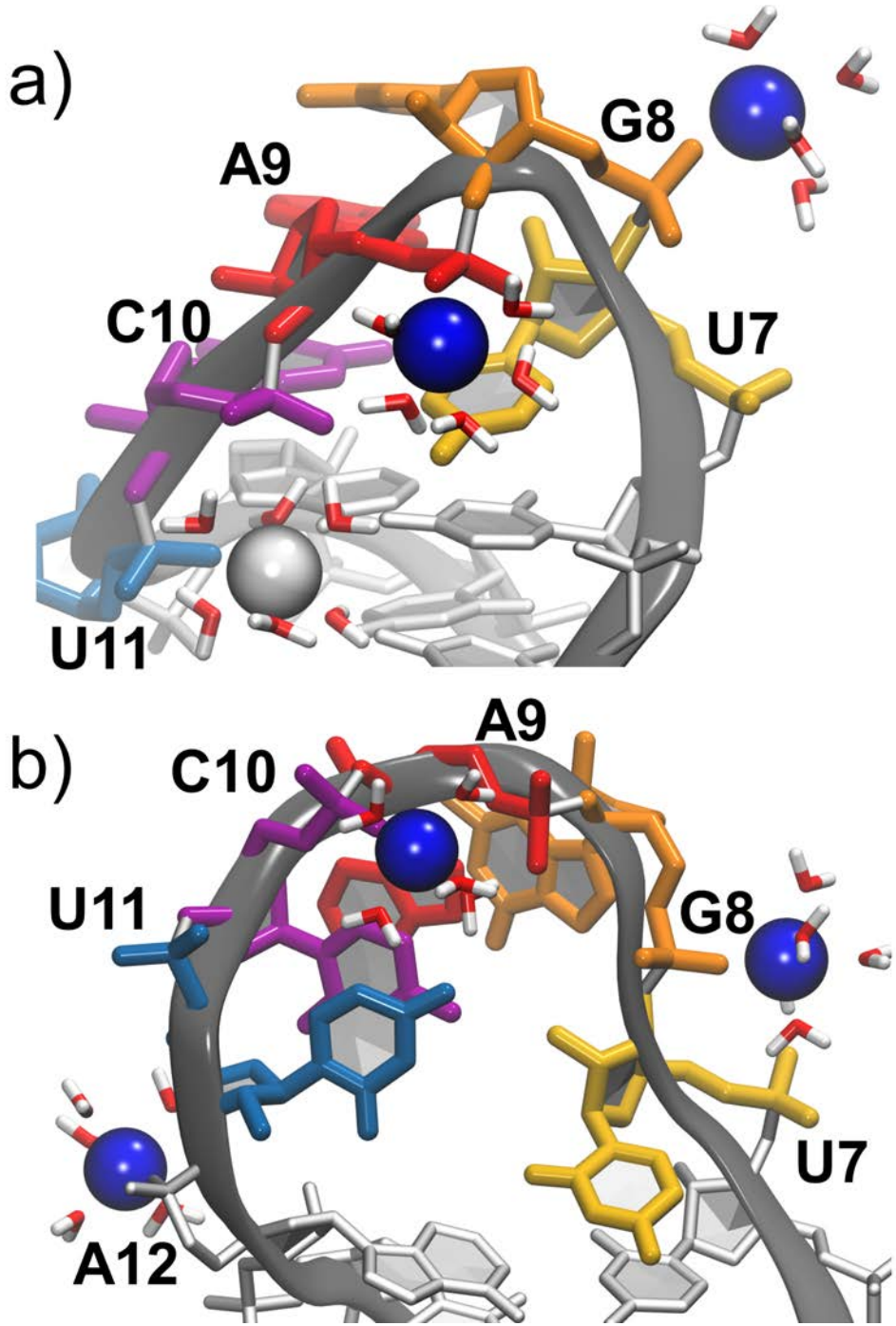
OK ☺



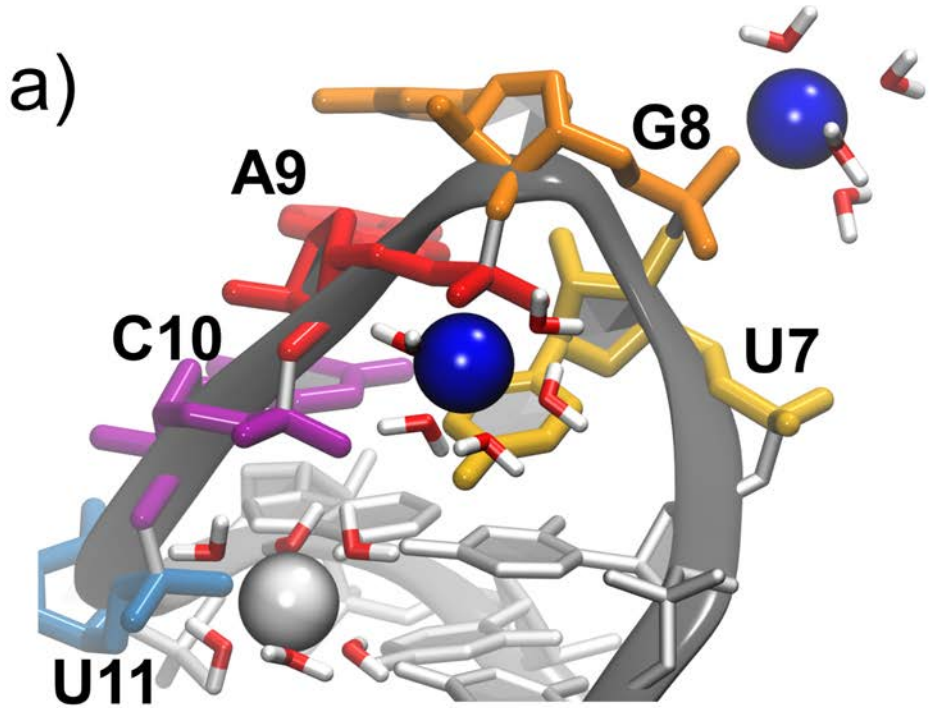
12-6-4 successful transition

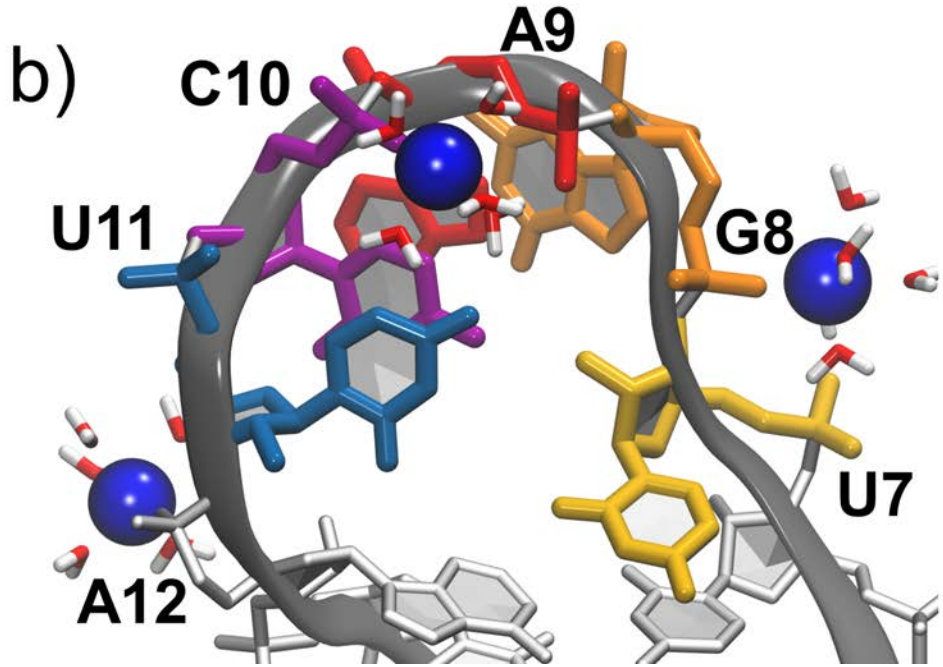
OK ☺

trapped for ms ☹



OK ☺





trapped for ms ☹

12-6-4 chelated ion affinity is 12-13.5 kcal/mol!

should the force field target the correct Mg^{2+} - water affinity?

What are some of the problems that still remain?

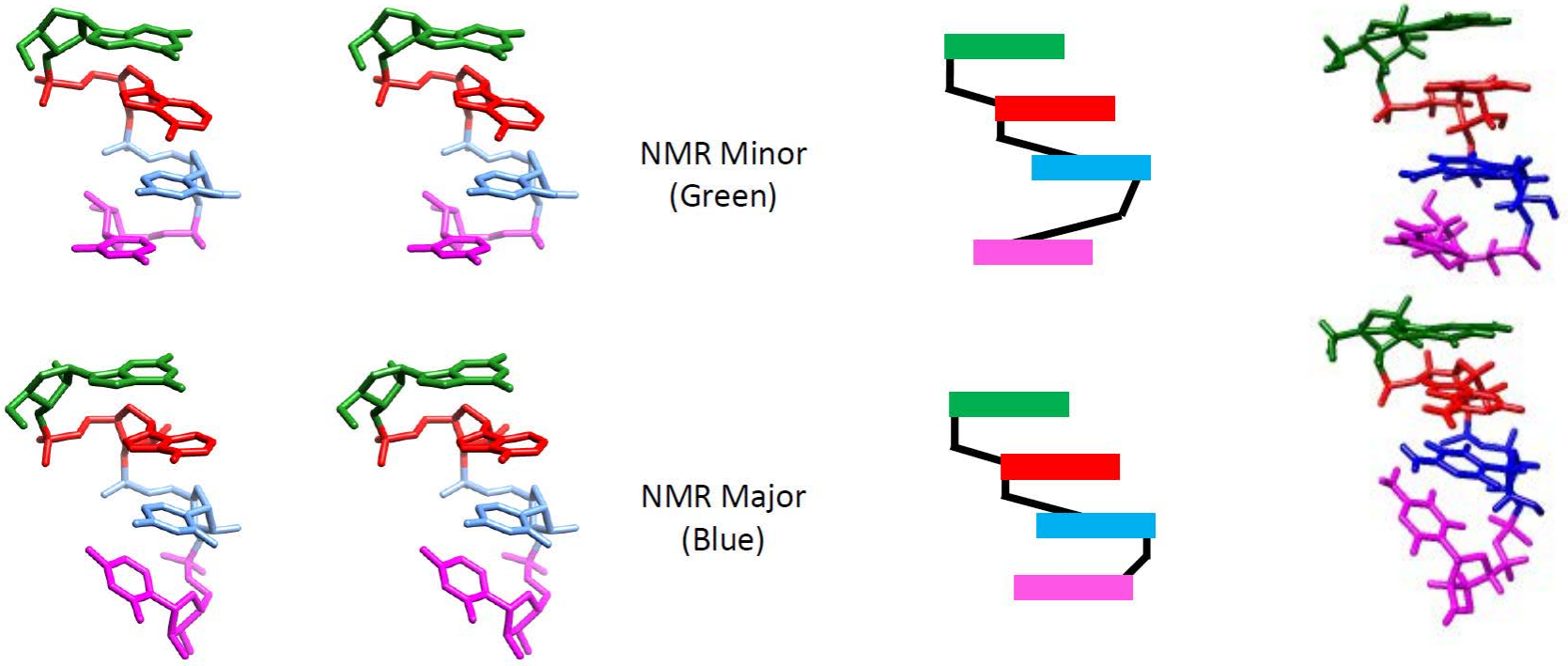
- Force field (mis)-balance
- Sampling

- [force field / methods inter-operability]
- [Judging convergence or overlap of independent simulations from different groups]

…a system where we can get complete sampling

r(GACC) tetranucleotide

[Turner / Yildirim]

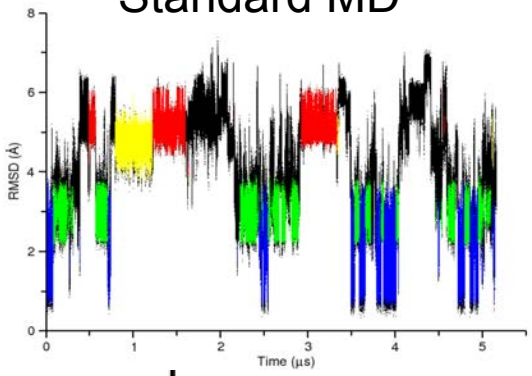


NMR suggests two dominant conformations...

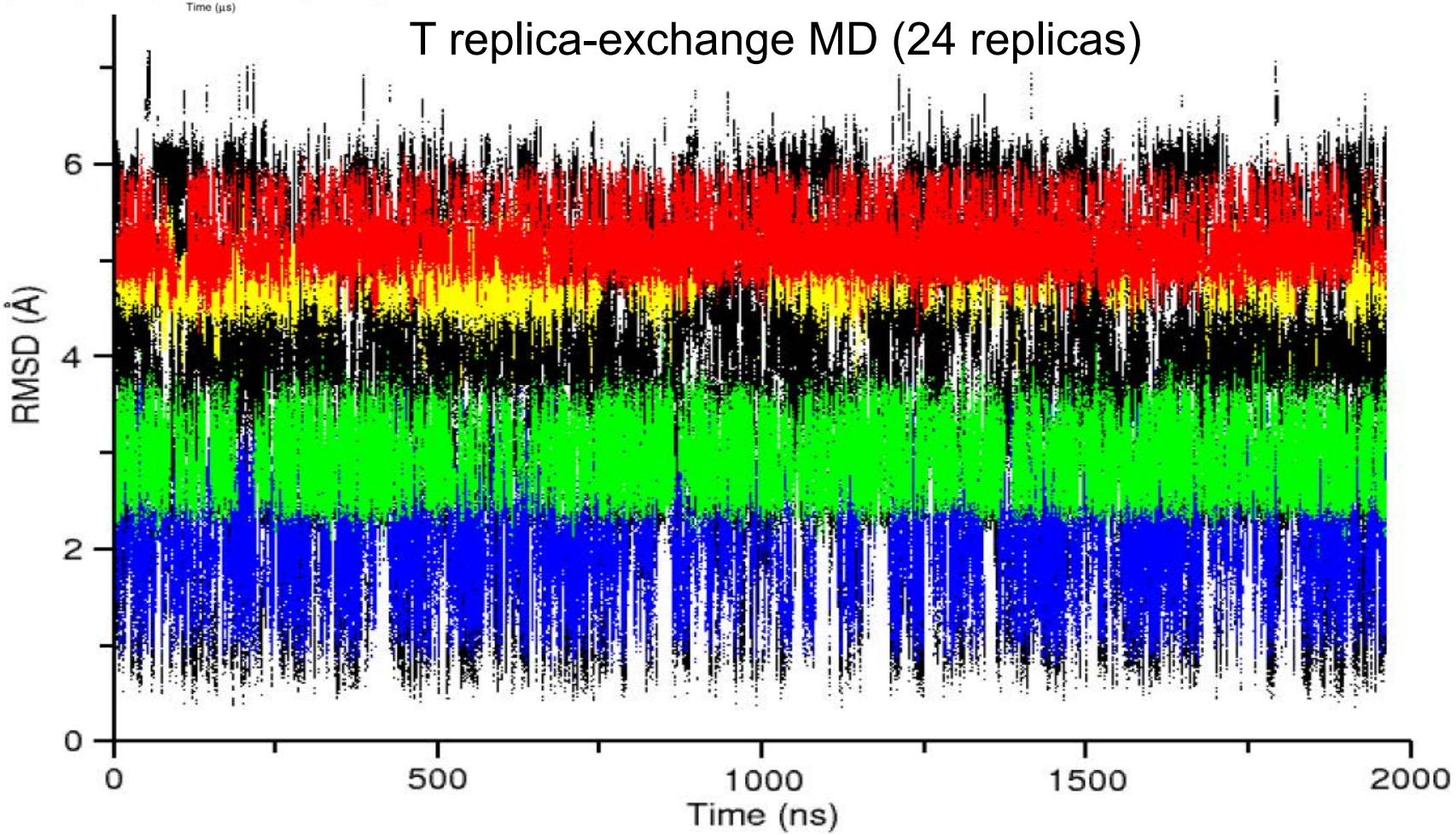
...compare to MD simulations in explicit solvent

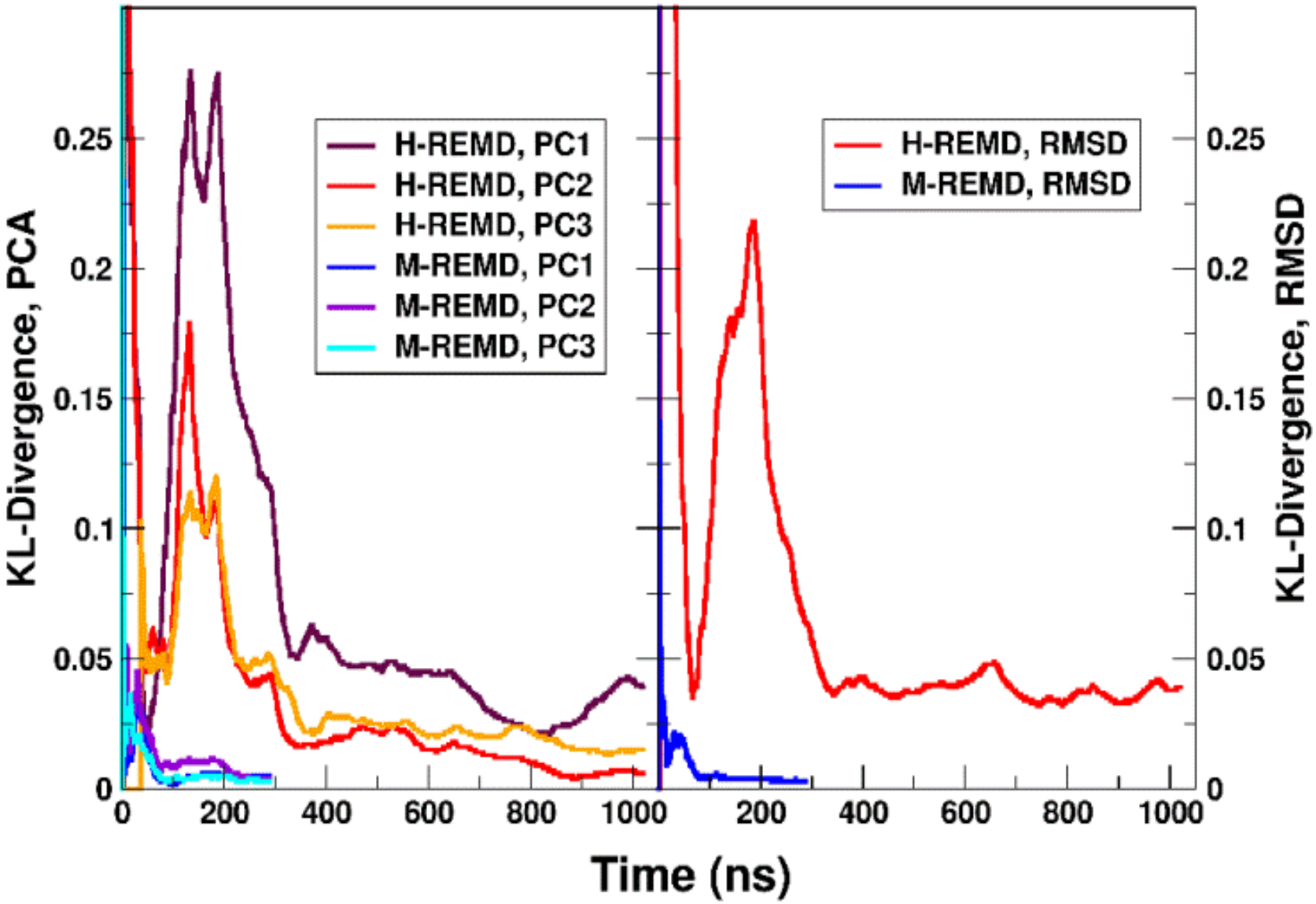
r(GACC) tetranucleotide

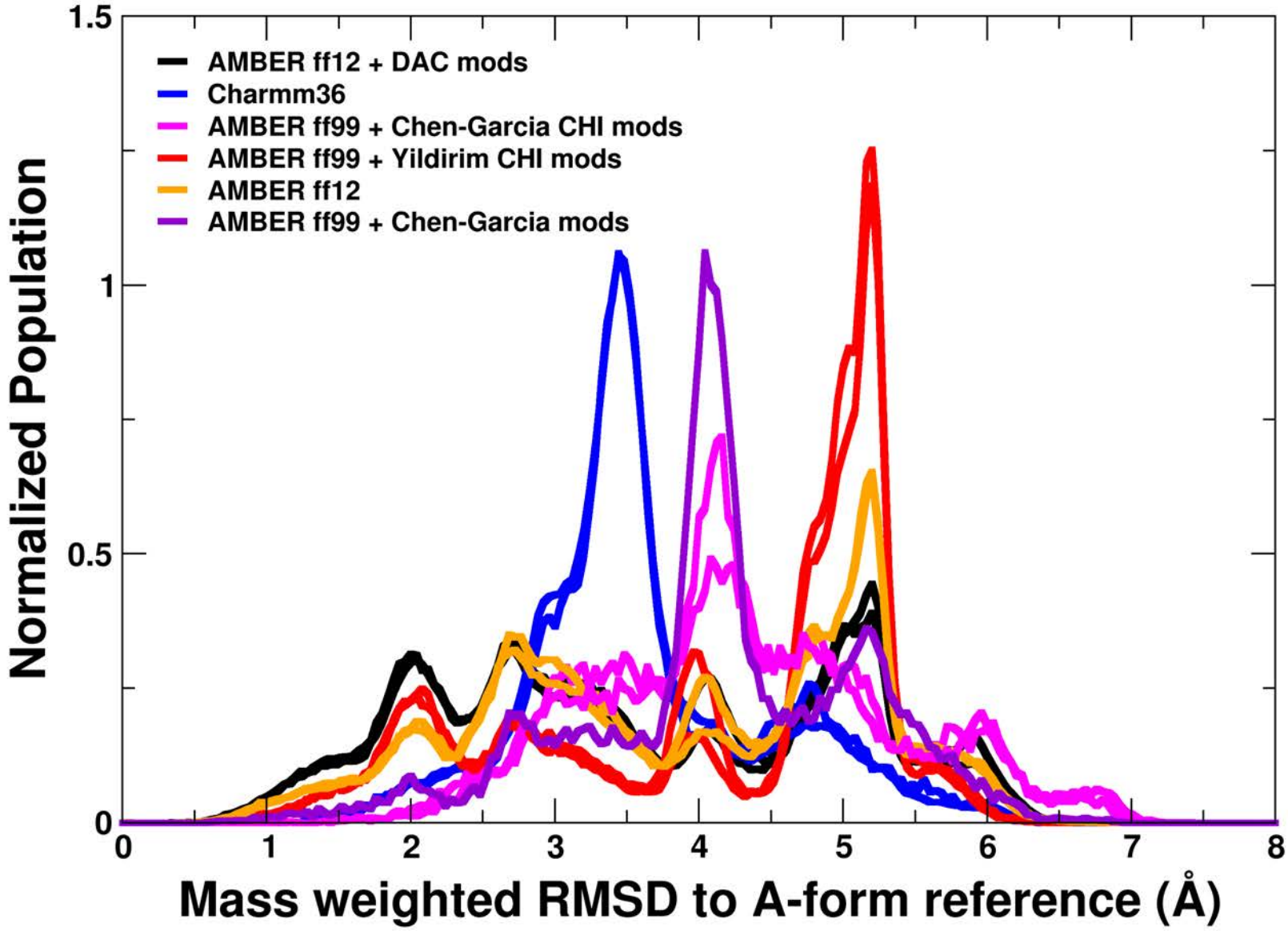
Standard MD

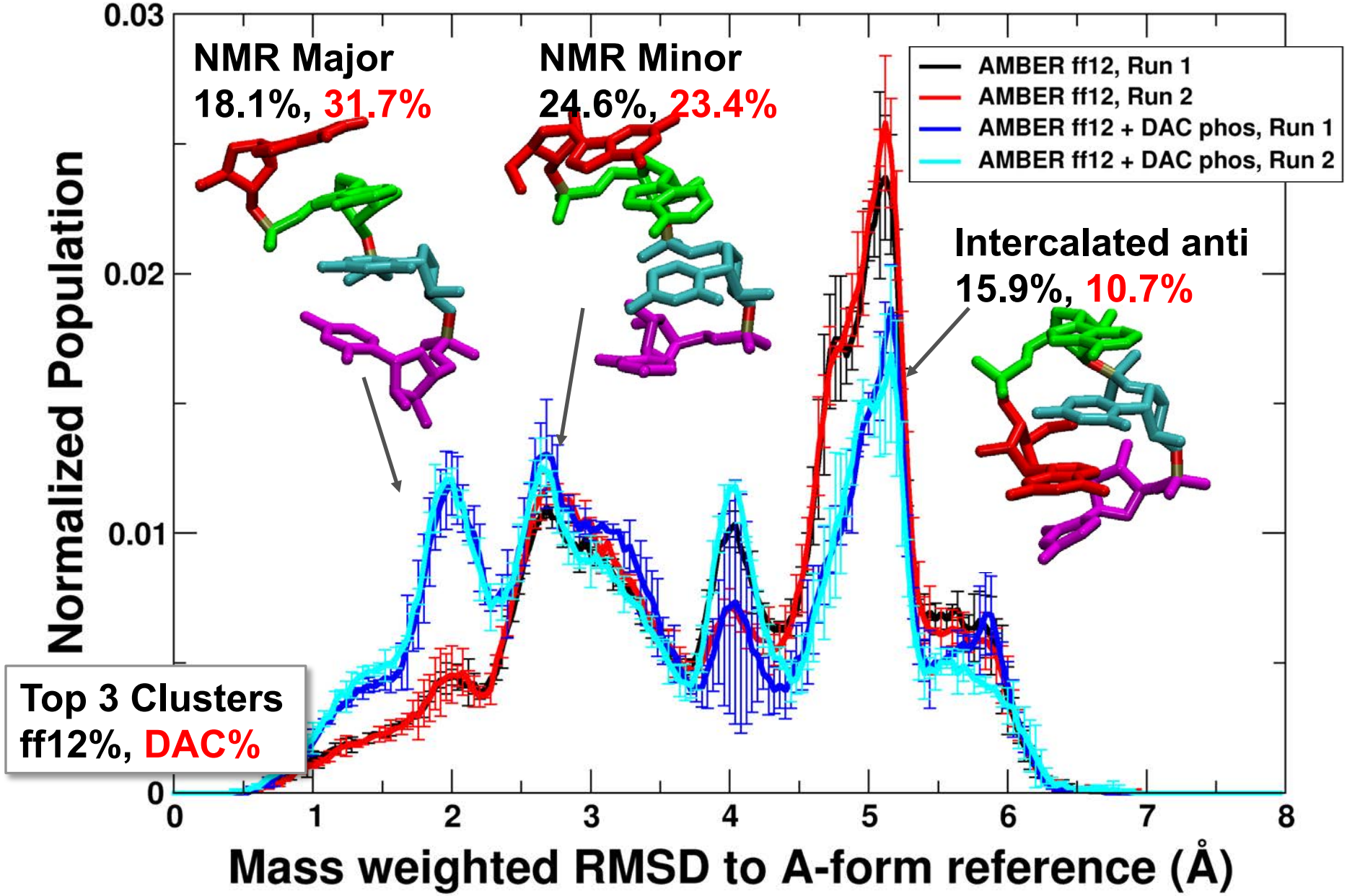


T replica-exchange MD (24 replicas)

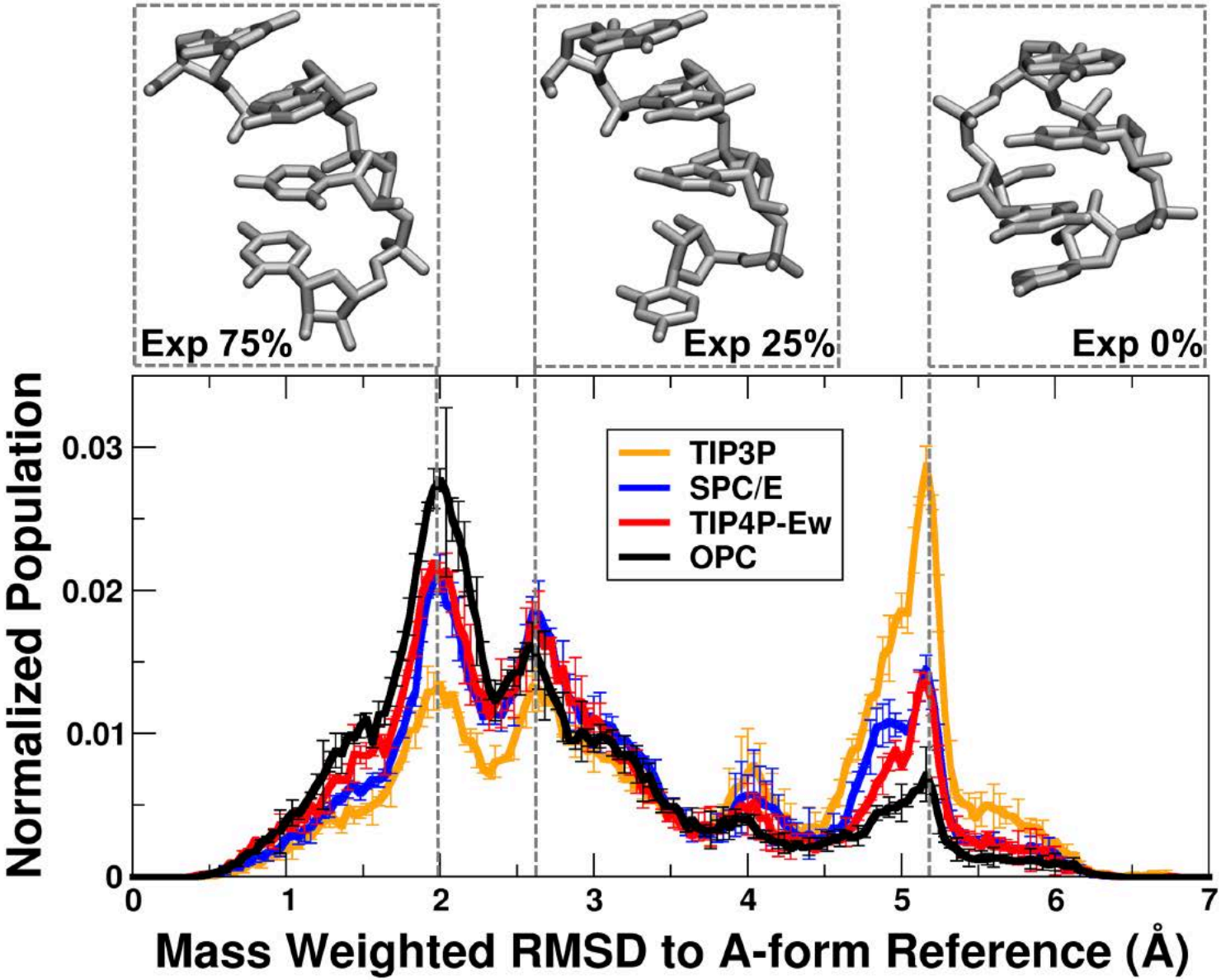


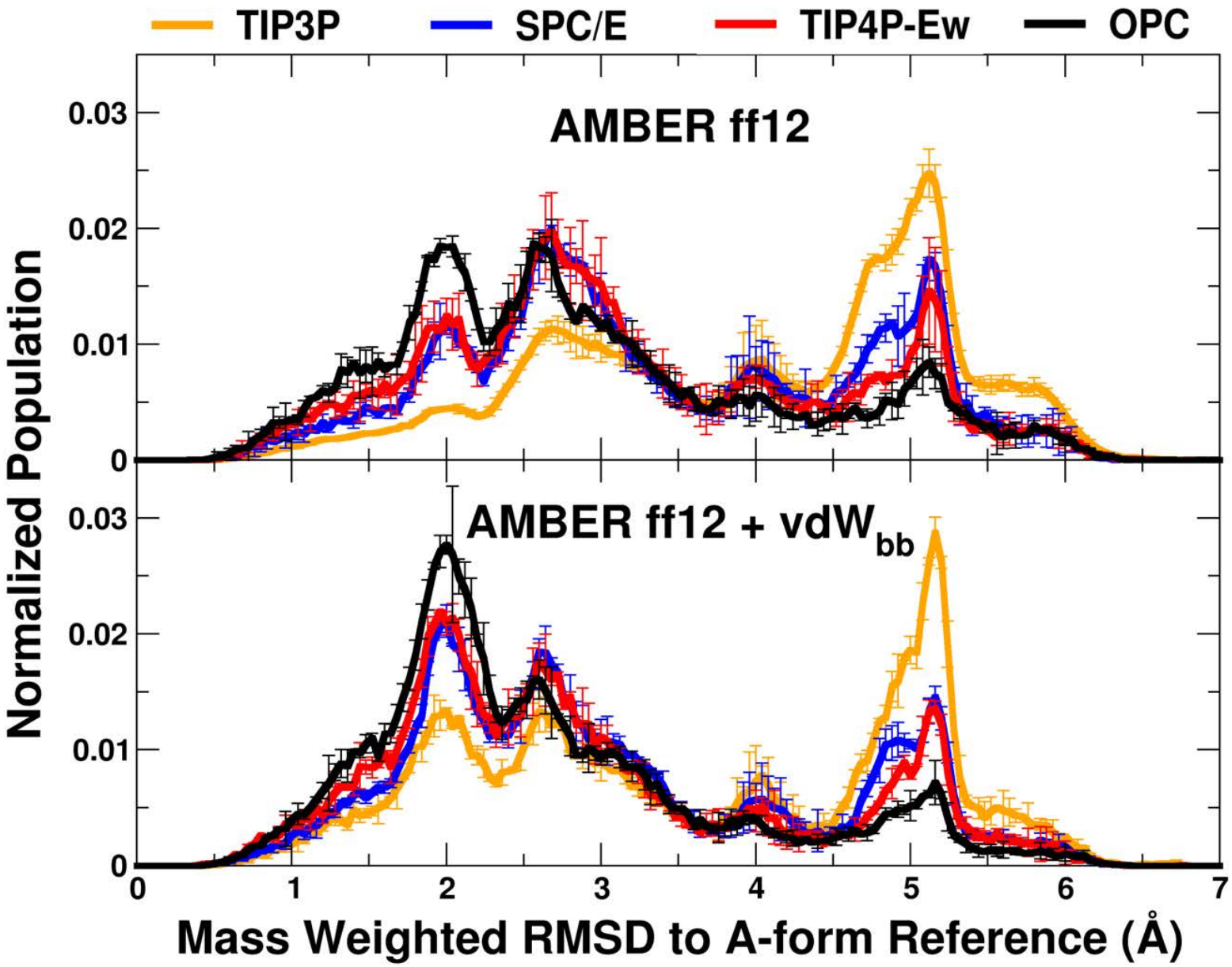

H-REMD, PC1
H-REMD, PC2
H-REMD, PC3
M-REMD, PC1
M-REMD, PC2
M-REMD, PC3
H-REMD, RMSD
M-REMD, RMSD
KL-Divergence, PCA
KL-Divergence, RMSD
Time (ns)
0
0.05
0.1
0.15
0.2
0.25
200
400
600
800
1000


GACC Ensemble, Force Field Comparison
AMBER ff12 + DAC mods
Charmm36
AMBER ff99 + Chen-Garcia CHI mods
AMBER ff99 + Yildirim CHI mods
AMBER ff12
AMBER ff99 + Chen-Garcia mods
Normalized Population
1.5
1
0.5
0
0
1
2
3
4
5
6
7
8
Mass weighted RMSD to A-form reference (Å)


Convergence Analysis, GACC Ensemble
1st half vs. 2nd half, Force Field Comparison
NMR Major
18.1%, 31.7%
NMR Minor
24.6%, 23.4%
Intercalated anti
15.9%, 10.7%
AMBER ff12, Run 1
AMBER ff12, Run 2
AMBER ff12 + DAC phos, Run 1
AMBER ff12 + DAC phos, Run 2
Top 3 Clusters
ff12%, DAC%
Normalized Population
0.03
0.02
0.01
0
0
1
2
3
4
5
6
7
8
Mass weighted RMSD to A-form reference (Å)

r(GACC) – change water model, alter phosphate size





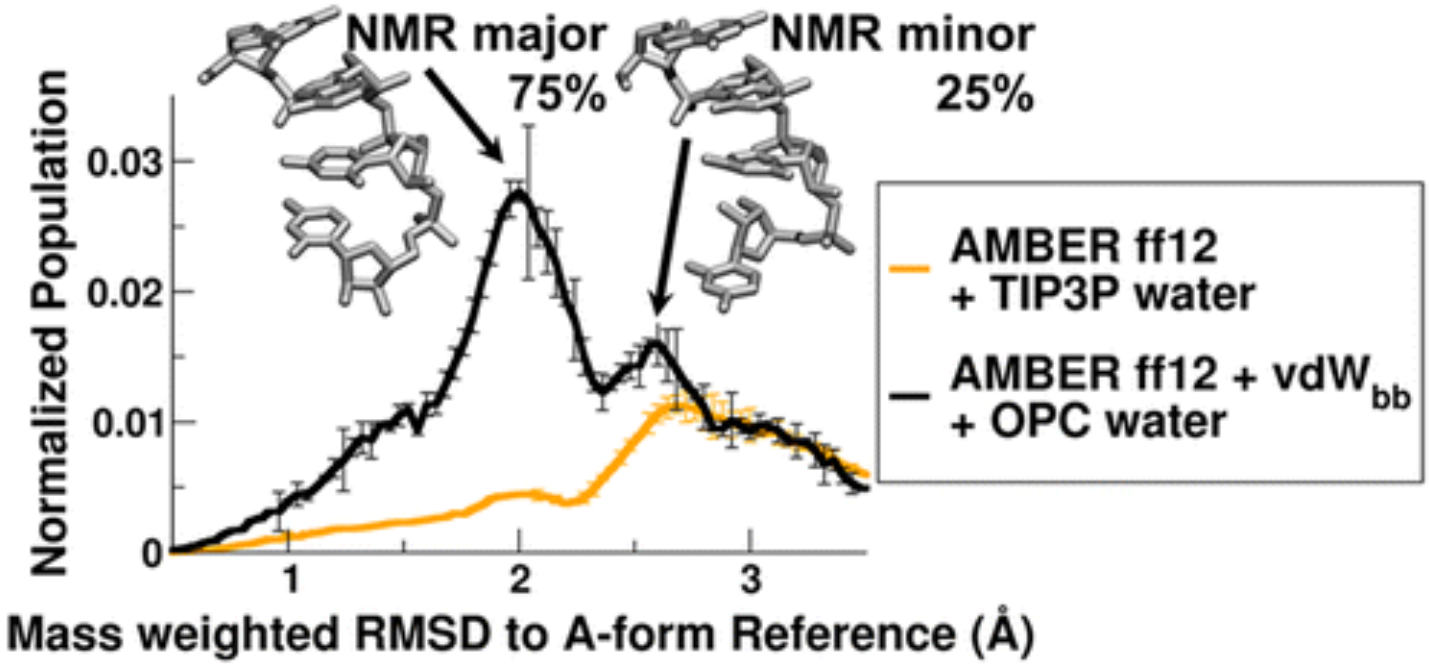
r(GACC): We now get correct 3:1 population of experimental structures with anomalous structures < 7%

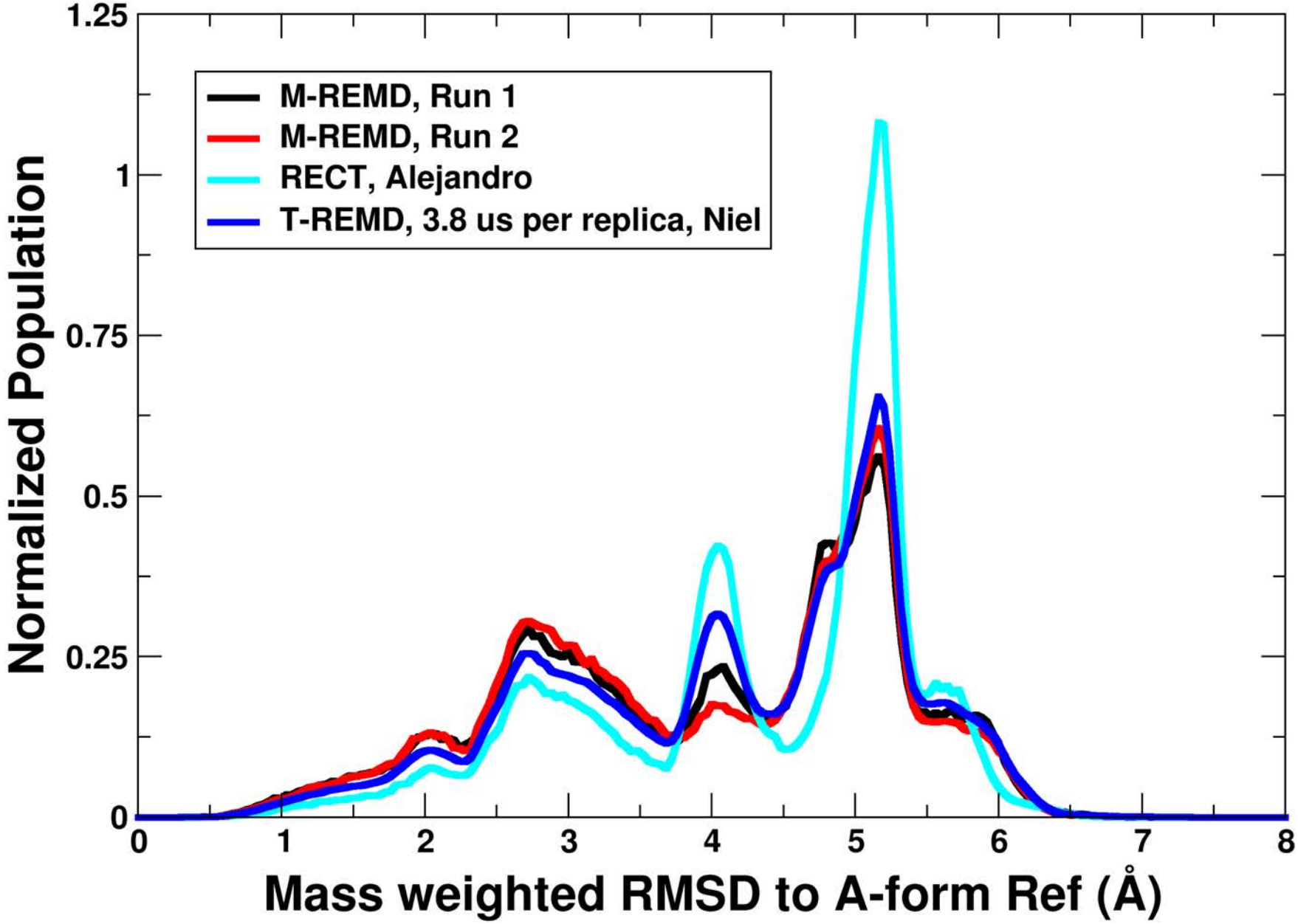
Letter

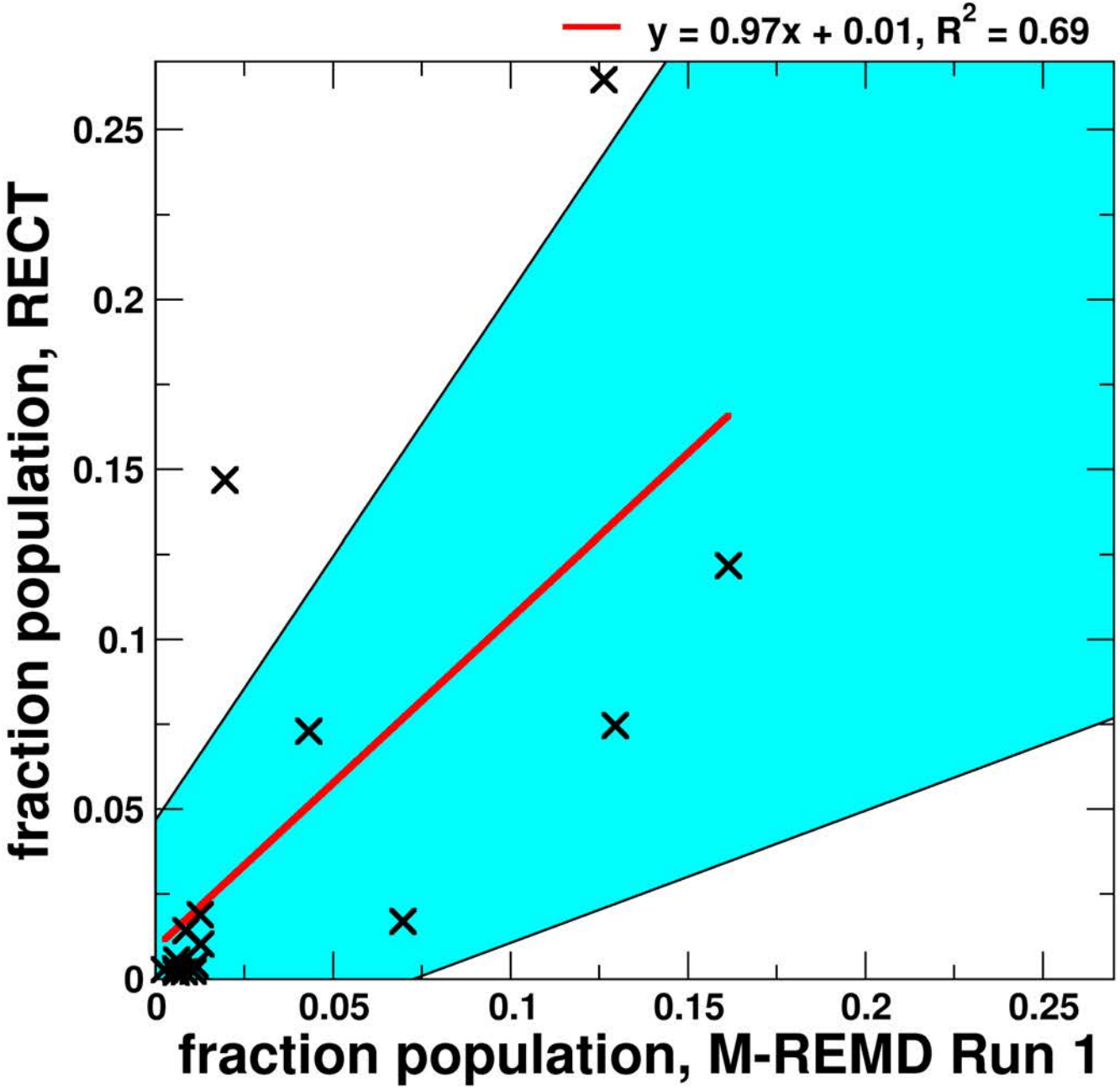
Improved Force Field Parameters Lead to a Better Description of RNA Structure

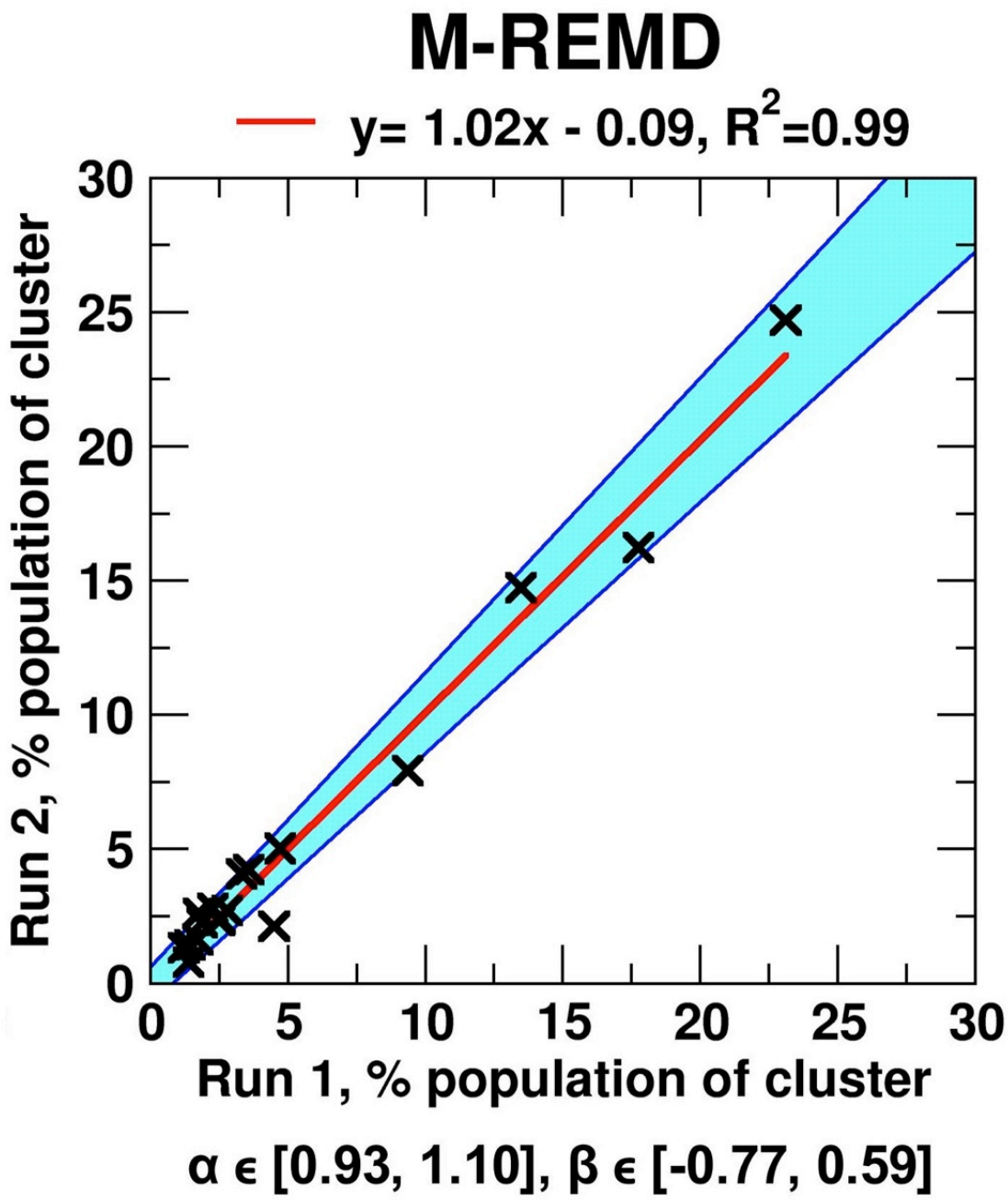
Christina Bergonzo and Thomas E. Cheatham III*
Department of Medicinal Chemistry, College of Pharmacy, University of Utah, Salt Lake City, Utah 84112, United States

J. Chem. Theory Comput., **2015**, *11* (9), pp 3969–3972
DOI: 10.1021/acs.jctc.5b00444
Publication Date (Web): August 7, 2015



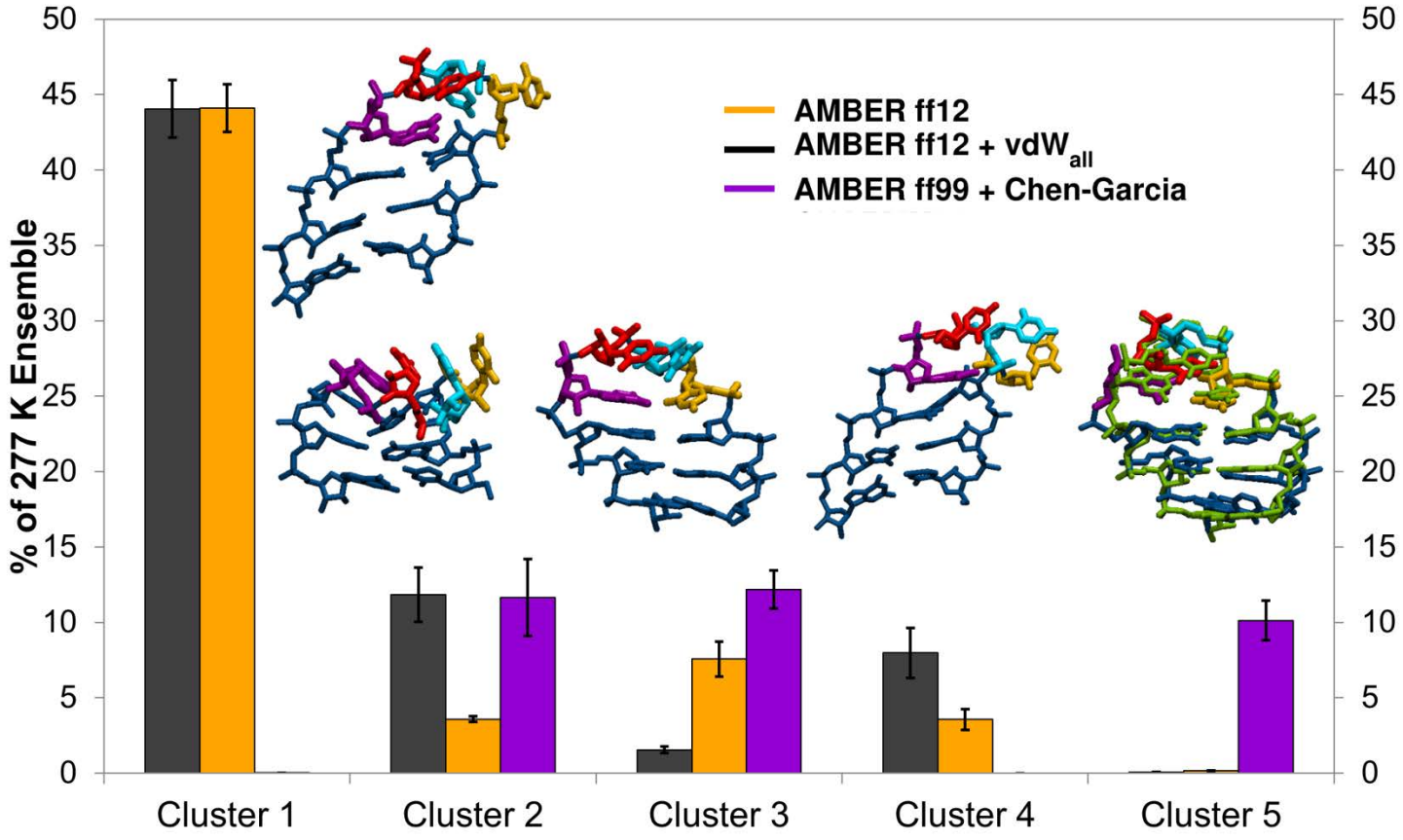

M-REMD, Run 1
M-REMD, Run 2
RECT, Alejandro
T-REMD, 3.8 us per replica, Niel
Normalized Population
Mass weighted RMSD to A-form Ref (Å)
0
0.25
0.5
0.75
1
1.25
0
1
2
3
4
5
6
7
8


y = 0.97x + 0.01, R2 = 0.69
fraction population, RECT
fraction population, M-REMD Run 1
0
0.05
0.1
0.15
0.2
0.25


M-REMD
y= 1.02x - 0.09, R2=0.99
Run 2, % population of cluster
Run 1, % population of cluster
0
5
10
15
20
25
30
α ϵ [0.93, 1.10], β ϵ [-0.77, 0.59]

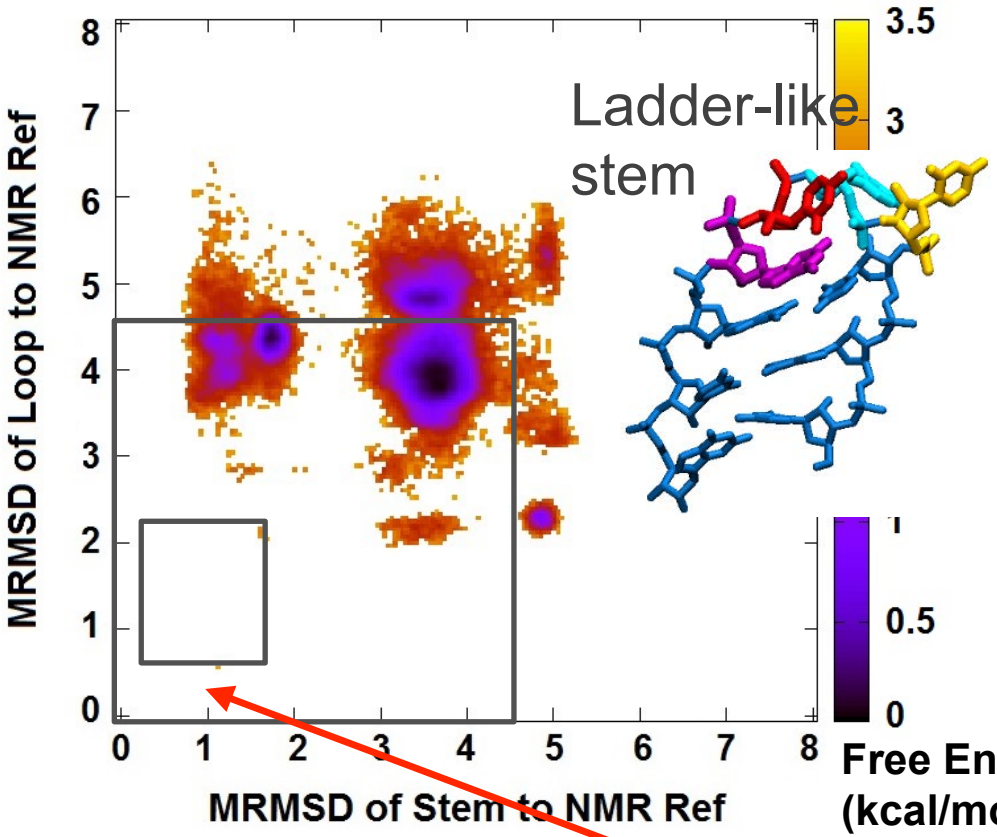
M-REMD can "converge" UUCG loop conformational ensemble

- folded UUCG tetraloop structure is sampled, but populations are low
- Chen-Garcia shifts distribution, but iso-energetic structures

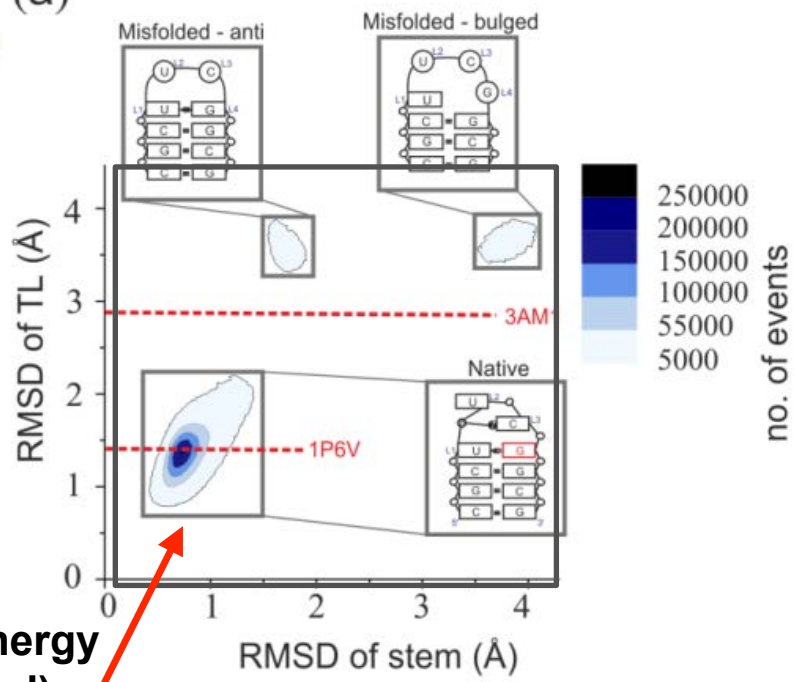


…more complete sampling may alter results

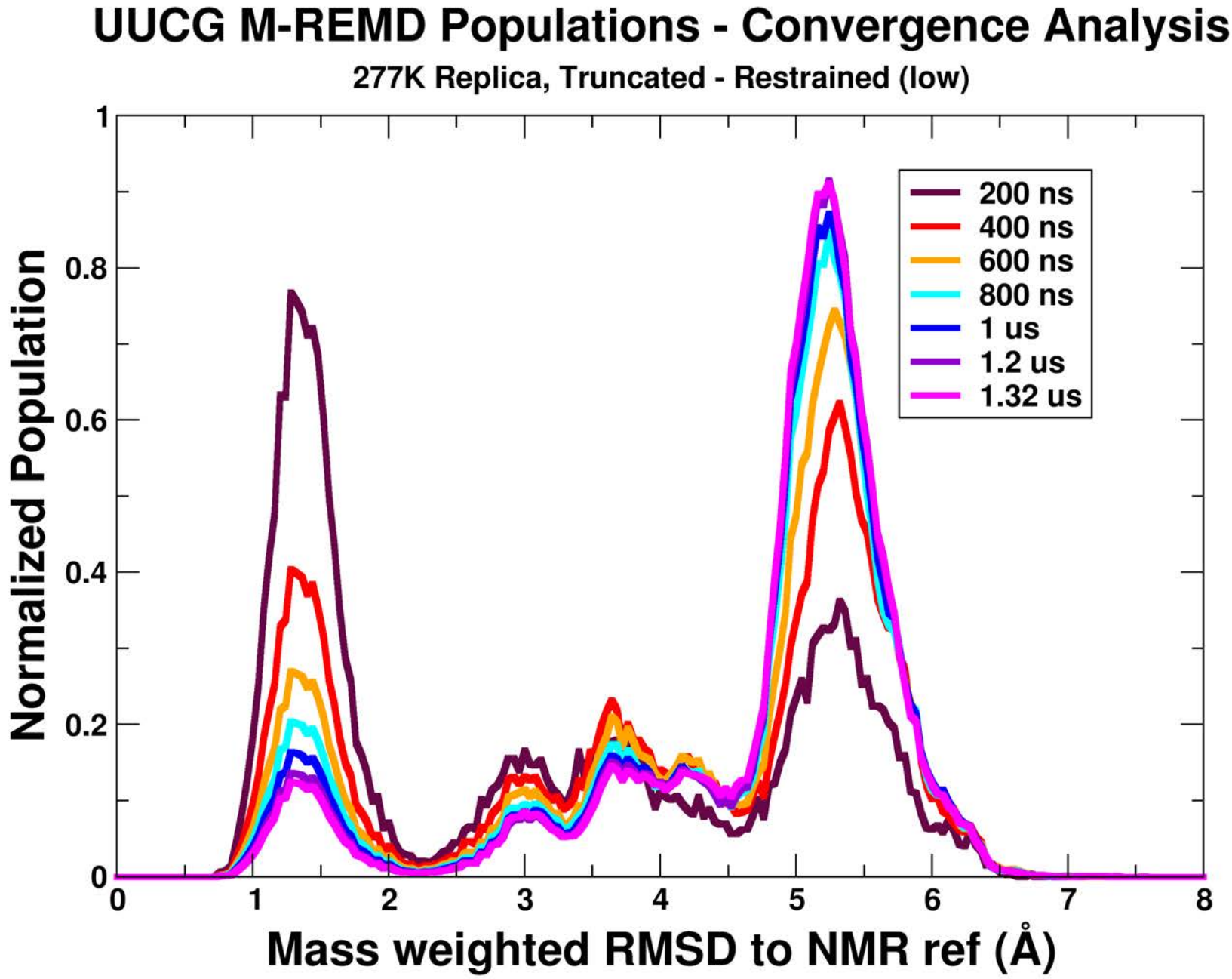
277 K – last 1 μs of 2 μs/replica M-REMD



Kührová et al. 2013 JCTC
500 ns T-REMD

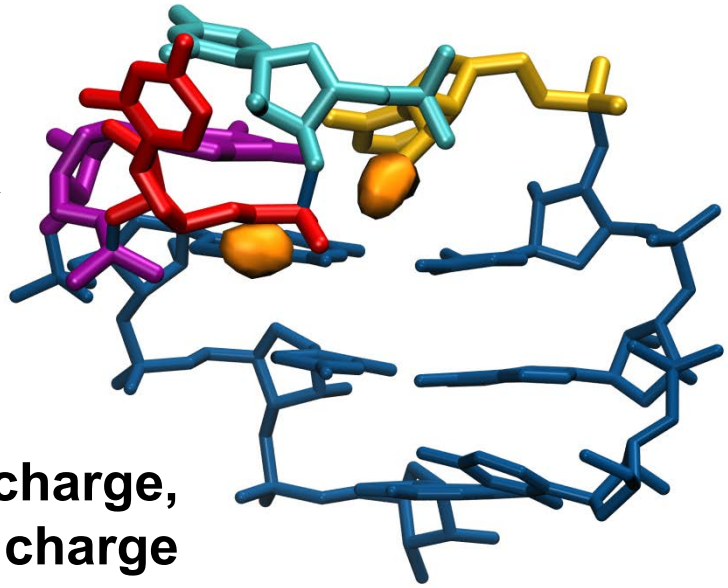


Native

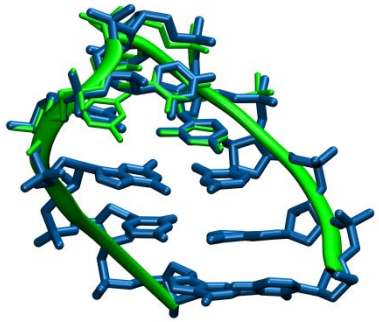

UUCG M-REMD Populations - Convergence Analysis
277K Replica, Truncated - Restrained (low)
Normalized Population
Mass weighted RMSD to NMR ref (Å)
200 ns
400 ns
600 ns
800 ns
1 us
1.2 us
1.32 us
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5
6
7
8

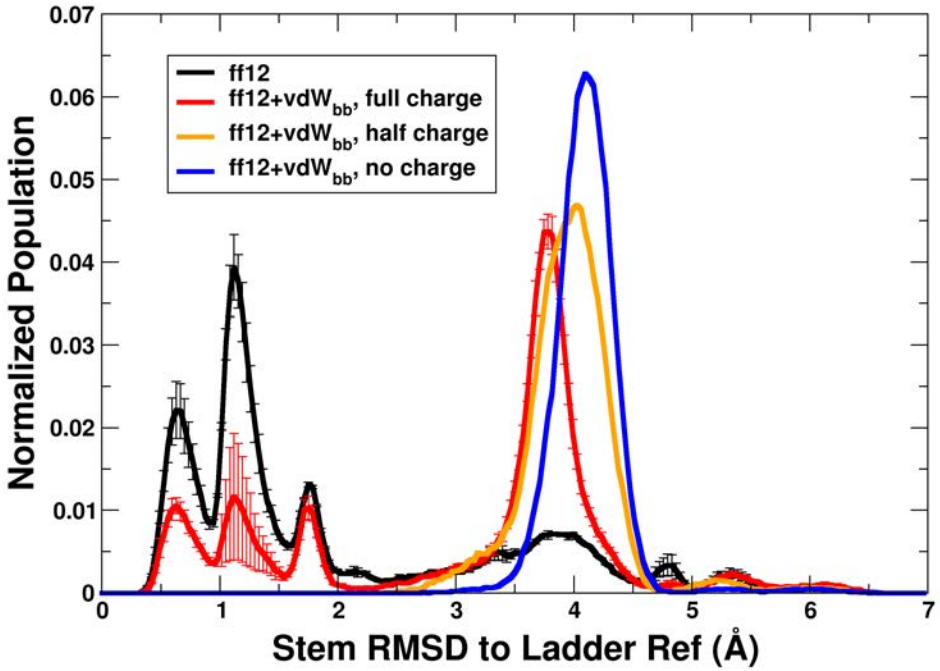
ff12+vdW$_{bb}$	Native*
% No Charge	17.7
% Half Charge	9.8
% Full Charge	0.41

Top Cluster

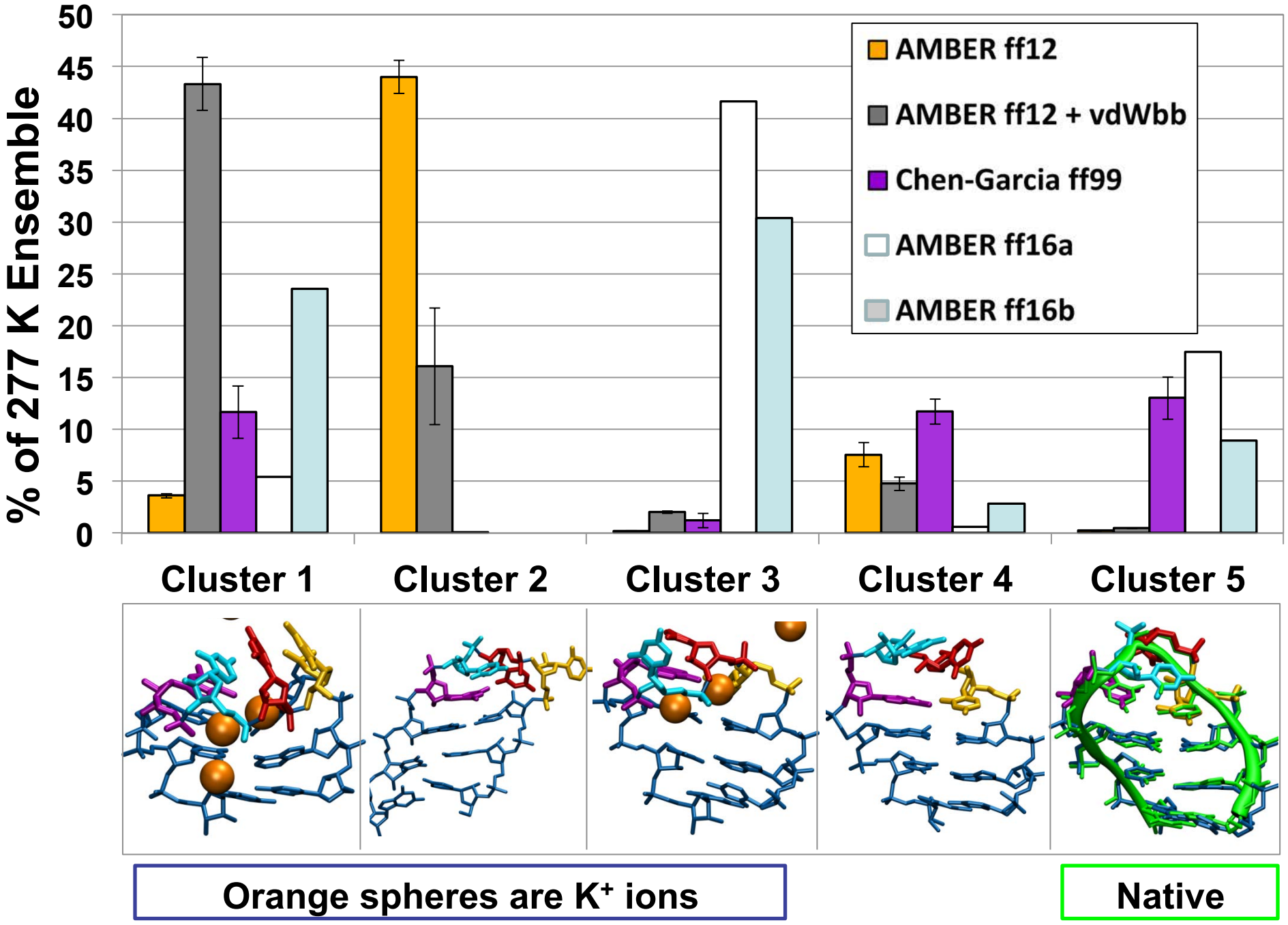


41.8% no charge,
55.7% half charge





Residue and Atom	**Fraction Occupancy, K+**	
	No Charge	Half Charge
C_6@OP2	1.70	1.05
G_8@O6	1.46	1.51
G_7@O6	0.93	0.91
G_8@N7	0.85	0.78
U_4@O2'	0.80	0.42
G_7@N7	0.79	0.77
G_7@OP2	0.78	0.48
U_5@OP2	0.43	0.46


% of 277 K Ensemble
50
45
40
35
30
25
20
15
10
5
0
AMBER ff12
AMBER ff12 + vdWbb
Chen-Garcia ff99
AMBER ff16a
AMBER ff16b
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Orange spheres are K+ ions
Native

2013
JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

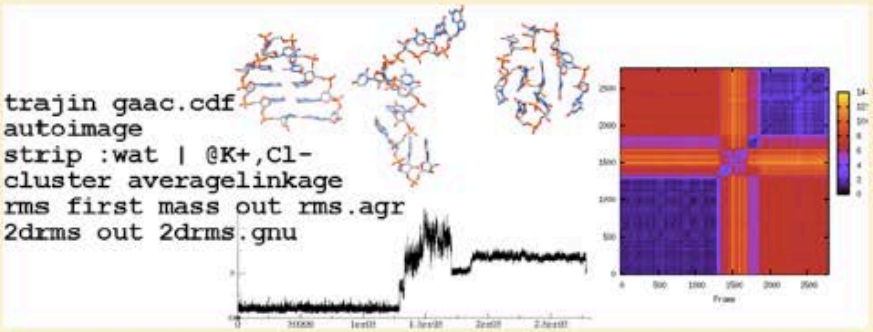
PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data

Daniel R. Roe* and Thomas E. Cheatham, III*

Department of Medicinal Chemistry, College of Pharmacy, 2000 South 30 East Room 105, University of Utah, Salt Lake City, Utah 84112, United States

Supporting Information

ABSTRACT: We describe PTRAJ and its successor CPPTRAJ, two complementary, portable, and freely available computer programs for the analysis and processing of time series of three-dimensional atomic positions (i.e., coordinate trajectories) and the data therein derived. Common tools include the ability to manipulate the data to convert among trajectory formats, process groups of trajectories generated with ensemble methods (e.g., replica exchange molecular dynamics), image with periodic boundary conditions, create average structures, strip subsets of the system, and perform calculations such as RMS fitting, measuring distances, B-factors, radii of gyration, radial distribution functions, and time correlations, among other actions and analyses. Both the PTRAJ and CPPTRAJ programs and source code are freely available under the GNU General Public License version 3 and are currently distributed within the AmberTools 12 suite of support programs that make up part of the Amber package of computer programs (see http://ambermd.org). This overview describes the general design, features, and history of these two programs, as well as algorithmic improvements and new features available in CPPTRAJ.



CUDA-CHiLL and CPPTRAJ

CHiLL/CUDA-CHiLL

- a source-to-source autotuning compiler (Hall, Utah)
- generates CUDA from sequential C/C++
- previously focused on DOE applications (SciDAC SUPER)

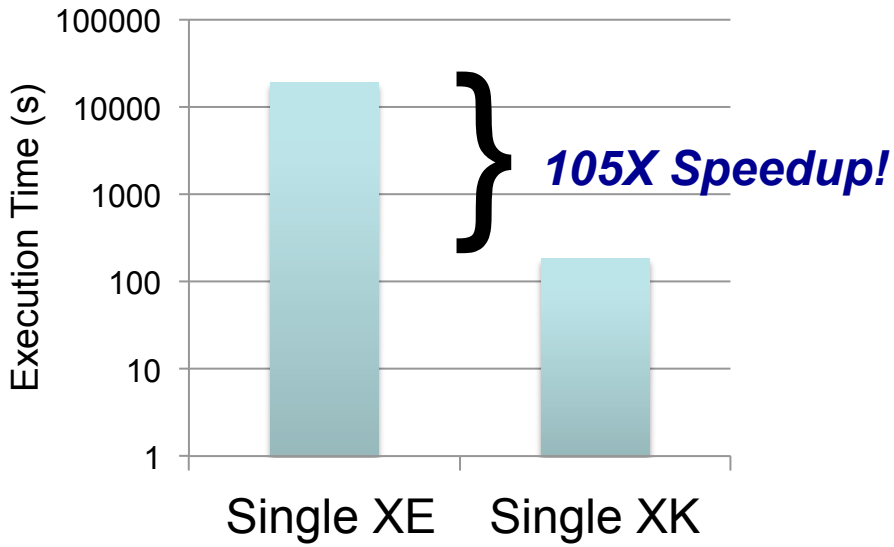
CPPTRAJ

- analysis code that examines results of simulation
- previously, MPI+OpenMP
- OpenMP threads used within actions
- scalability limited to small number of OMP threads

Improvement Method Enabler (IME) Project

Use CUDA-CHiLL to generate optimized CPPTRAJ GPU code on Blue Waters

CPPTRAJ Performance Results Example



CPPTRAJ Details

- Determine 965 closest solvent molecules out of 15022 to 4143 solute atoms, 2000 frames.
- Distance using periodic imaging.

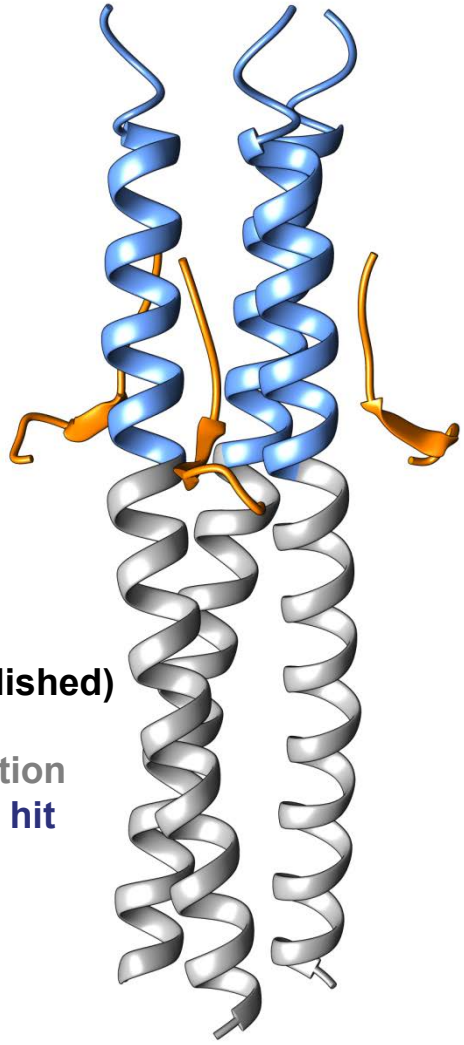
CUDA-CHiLL Details

- Required manually copying structure-of-arrays to set of arrays (also better for GPU global memory coalescing)
- Iterator through structure also had to be manually rewritten
- Also encountered minor bugs and limitations
- Plan to integrate improvements from this exercise!

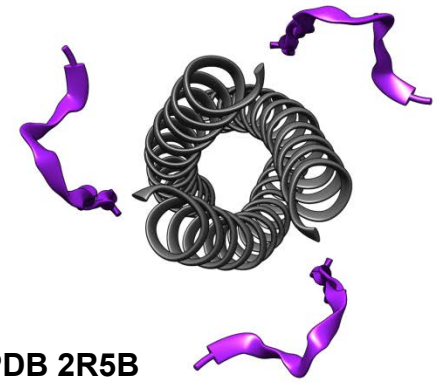
D-peptide Inhibitors of HIV & Ebolavirus Viral Entry

- Kay Lab (U of Utah)
- Short D-peptides block viral entry of HIV & ebolavirus.
- Utilize modified variant of phage display:
  - Random sequences presented to GP41(HIV) GP2(ebola).
- HIV = '**PIE7**' & '**PIE12**' D-peptides
- Ebola = '**HLLY6**' D-peptide
  - Each bind in 3:1 stoichiometry
- **HOWEVER**: workflow is extremely limited in effectiveness.
  - Leads = low micromolar binders
  - Lead opt. is lengthy & expensive process

Ebolavirus GP2 Mimic + HLLY6



HIV GP41 Mimic + PIE12

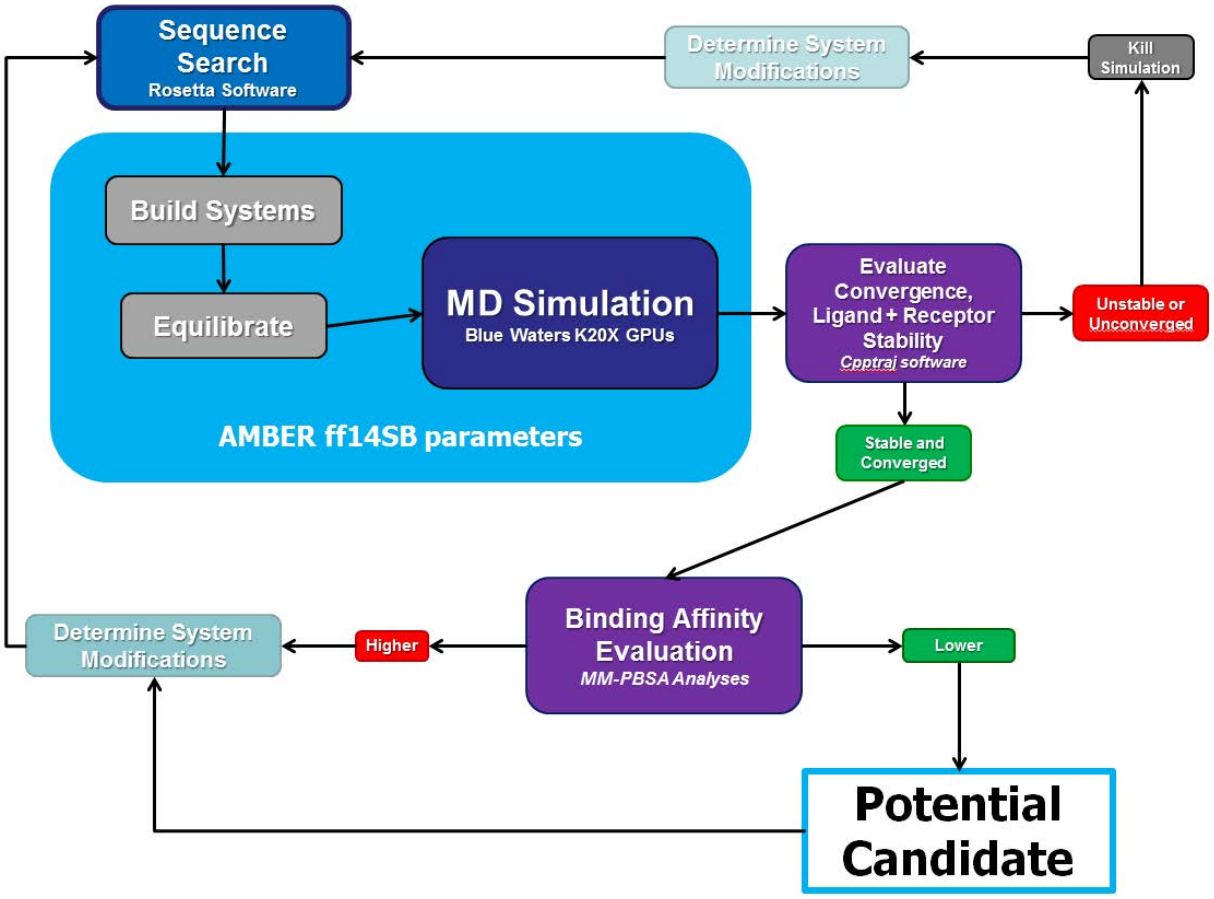


PDB 2R5B

- **Ebola GP2 + Initial 'hit' (unpublished)**
- **Blue region = GP2 Receptor**
- **Grey region = IZ_m anti-aggregation**
- **Orange region = HLLY6 11-mer hit**

See Sean's poster!

Novel Software Framework to do Heavy Lifting



General Workflow

1. Sequence Search
2. Build Systems
3. Min/Equil
4. Production MD
5. Structural Analysis
6. MM-PBSA

Integrating with Ensemble Toolkit / Radical Pilot (Jha @ Rutgers)

static vs. adaptable simulation / analysis patterns (Jha, Kasson, Shirts)

People: Niel Henriksen, Hamed Hayatshahi, Dan Roe, Julien Thibault, Kiu Shahrokh, Rodrigo Galindo, Christina Bergonzo, Sean Cornillie

$$$:

National Science Foundation
WHERE DISCOVERIES BEGIN

R01-GM098102:	"RNA-ligand interactions: sim. & experiment	~2015
R01-GM072049:	"P450 dehydrogenation mechanisms"	~2014
R01-GM081411:	"…simulation … refinement of nucleic acid"	~2013
NSF CHE-1266307 "CDS&E: Tools to facilitate deeper data analysis, …"		~2015
NSF "Blue Waters" PetaScale Resource Allocation for AMBER RNA		

Computer time:

"Anton"
(3 past awards)

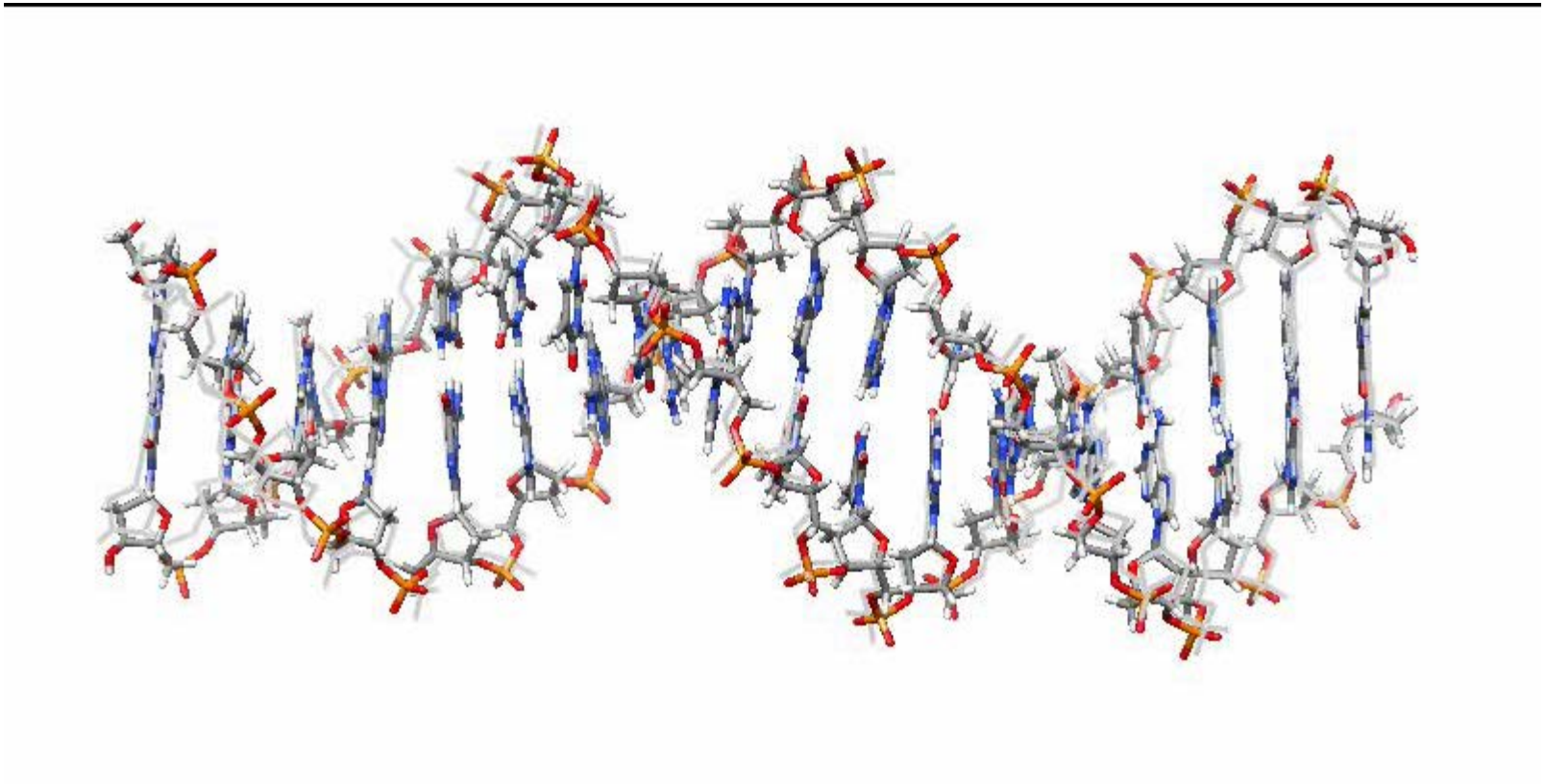
XRAC MCA01S027
~15M core hours

~14M GPU hours

!!!

~3M hours

questions?



2 ns intervals, 10 ns running average, every 5th frame (~10 us).