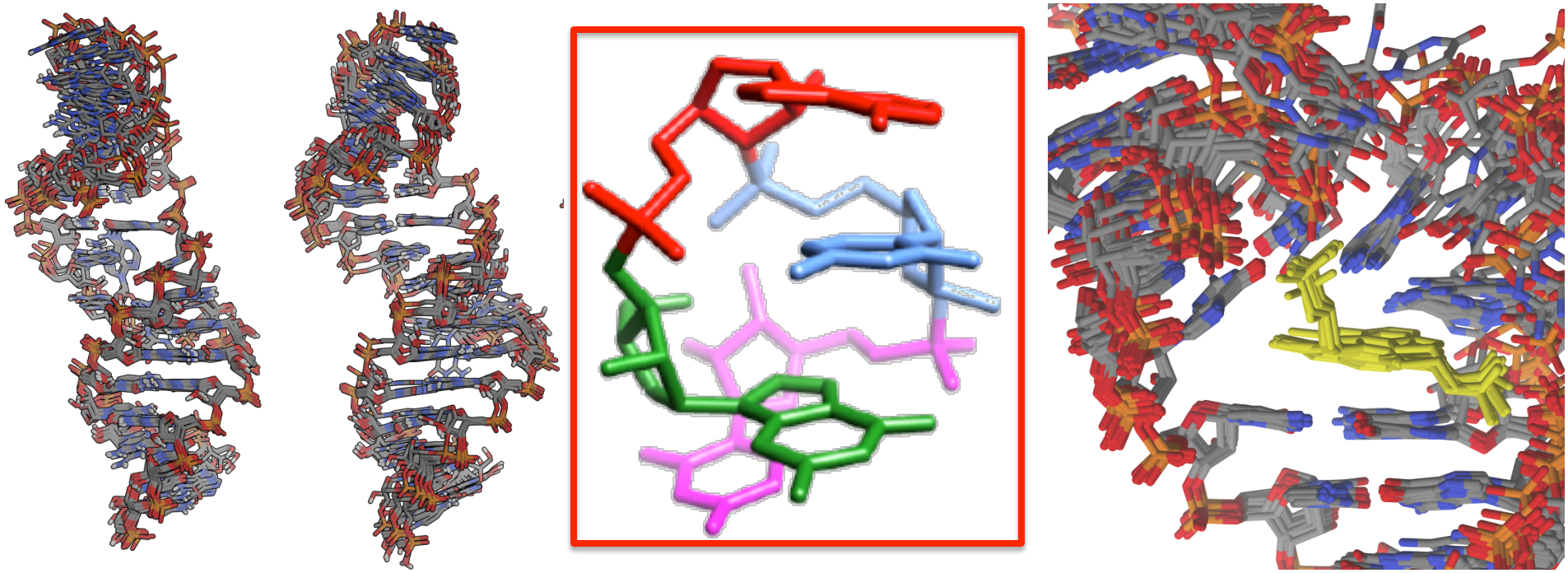
Using large-scale simulation to reproducibly converge nucleic acid structure and dynamics



Thomas E. Cheatham III
Associate Professor
Department of Medicinal Chemistry
College of Pharmacy, University of Utah

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:

PITTSBURGH SUPERCOMPUTING CENTER

D E Shaw Research

"Anton"
(3 past awards)

XSEDE
Extreme Science and Engineering Discovery Environment

XRAC MCA01S027
~15M core hours

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

~14M GPU hours

!!!

CHPC

~3M hours

Accurate modeling of RNA and other biomolecules requires:

accurate and fast simulation methods

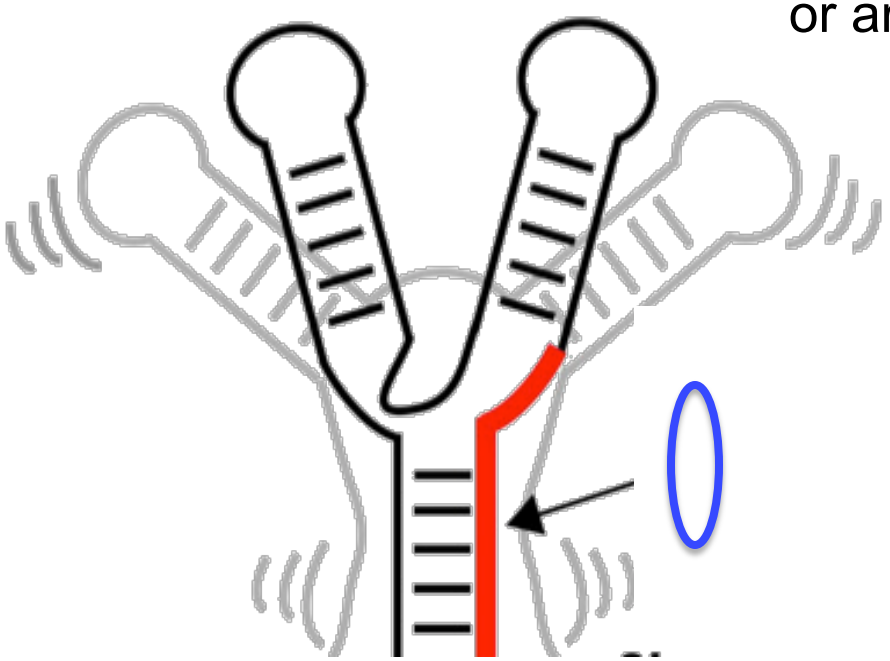
validated RNA, protein, water, ion, and ligand "force fields"

"good" experiments to assess results

dynamics and complete sampling: (convergence, reproducibility)

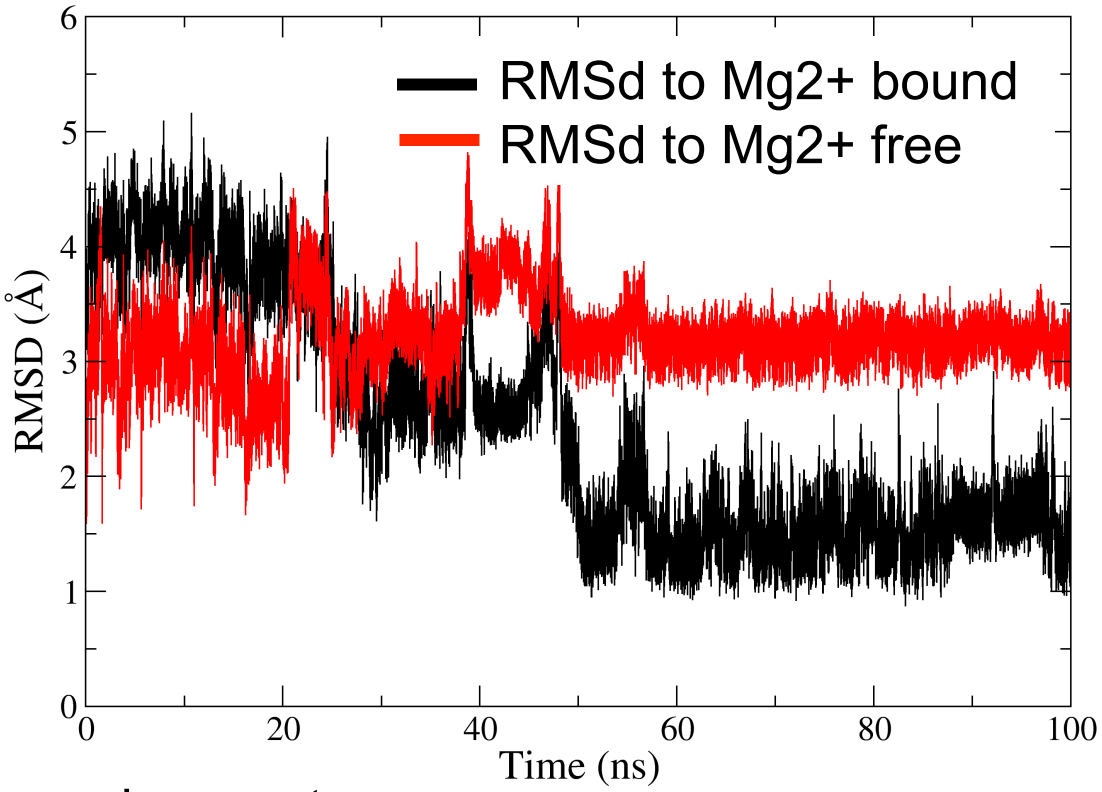
Question: Is the movement real or artifact?

True RNA dynamics
or artifact of force field?



We're seeing some progress!!!

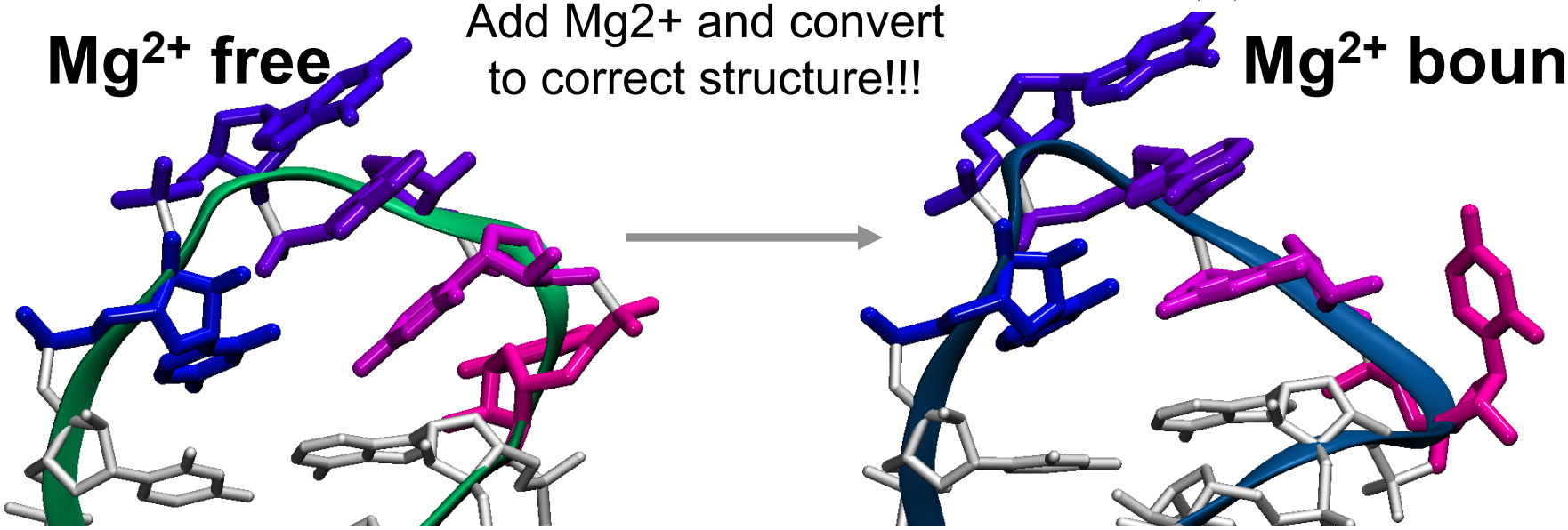
(vsrSL5)



Mg^{2+} free

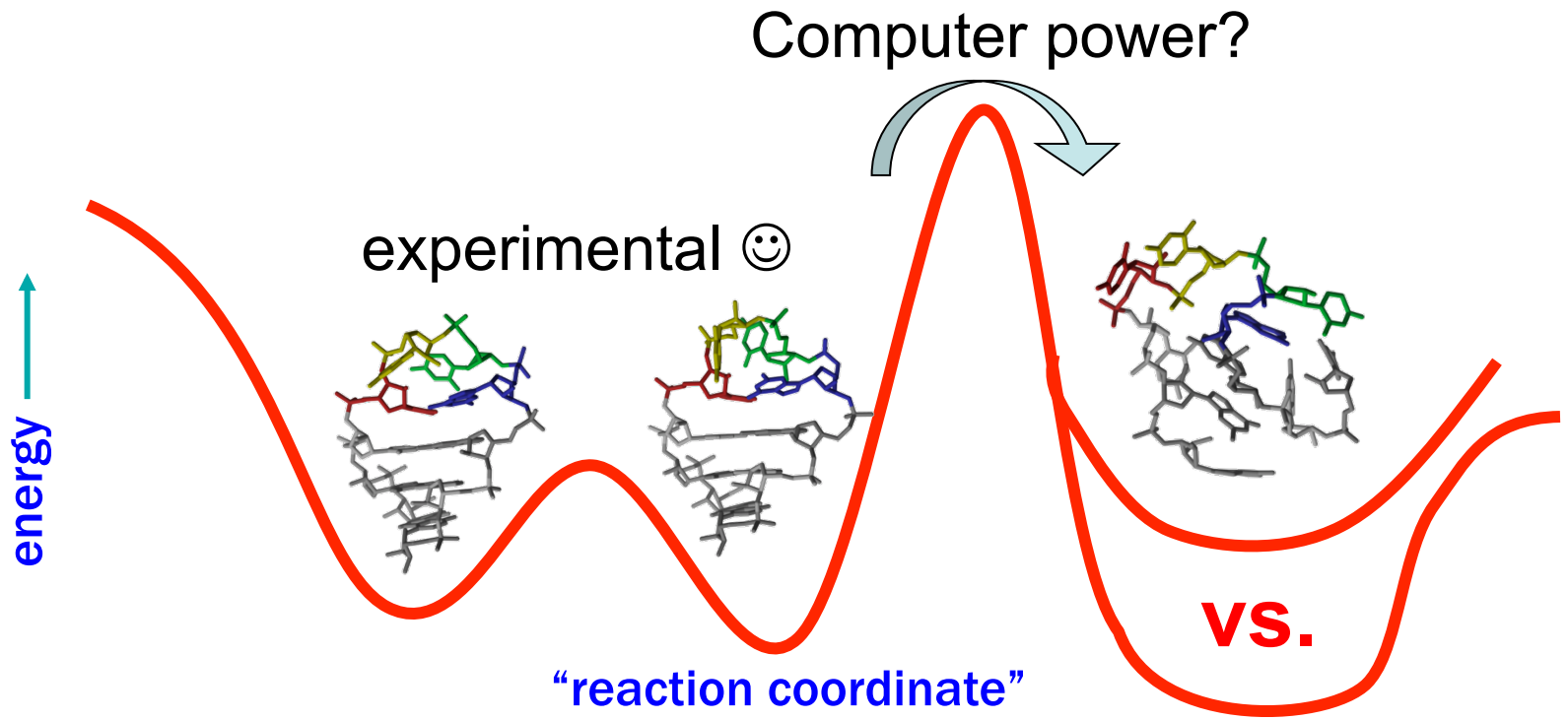
Add Mg2+ and convert to correct structure!!!

Mg^{2+} bound

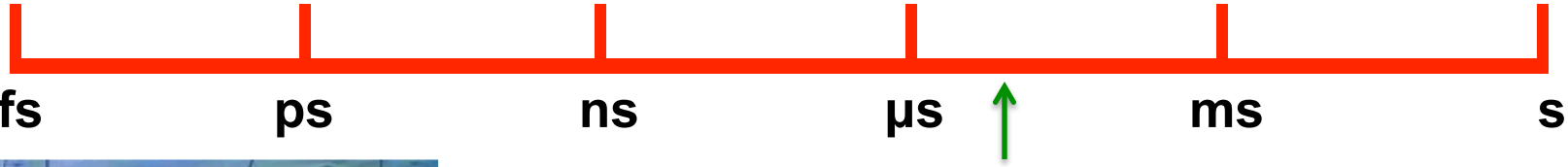


are the force fields reliable?
(free energetics, sampling, dynamics)

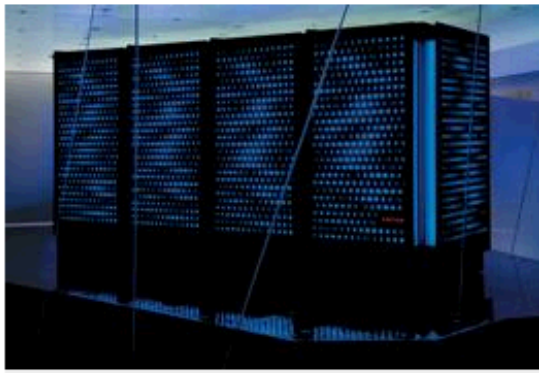
Short simulations stay near experimental structure; longer simulations invariably move away and often to unrealistic lower energy structures…



How to fully sample conformational ensemble?



16 μs/day!

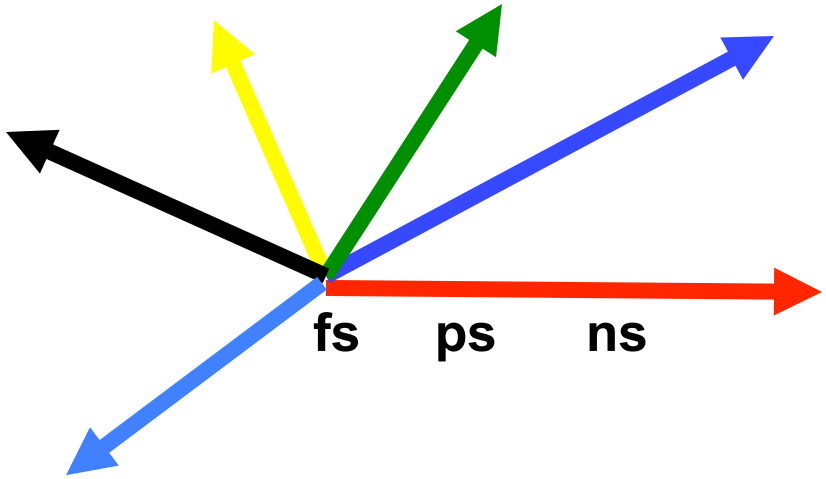


Simulating protein movements using Anton could aid drug design.

SCIENCE/AAAS

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

D.E. Shaw's Anton machine



ensembles of independent simulations

AMBER on GPUs

110 ns/day!

~1978 - present

amber

Assisted Model Building with Energy Refinement

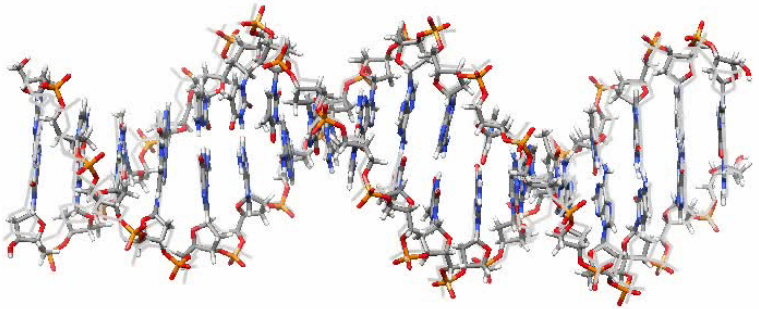
code vs. force field

Amber 14 released April, 2014

- 1.23x increase in GPU performance; peer-to-peer
  [fully deterministic, mixed SP/fixed precision, ||-ized]
- **support for M-REMD simulation and analysis**
- constant pH
- new TI methods
- more methods ported to GPU

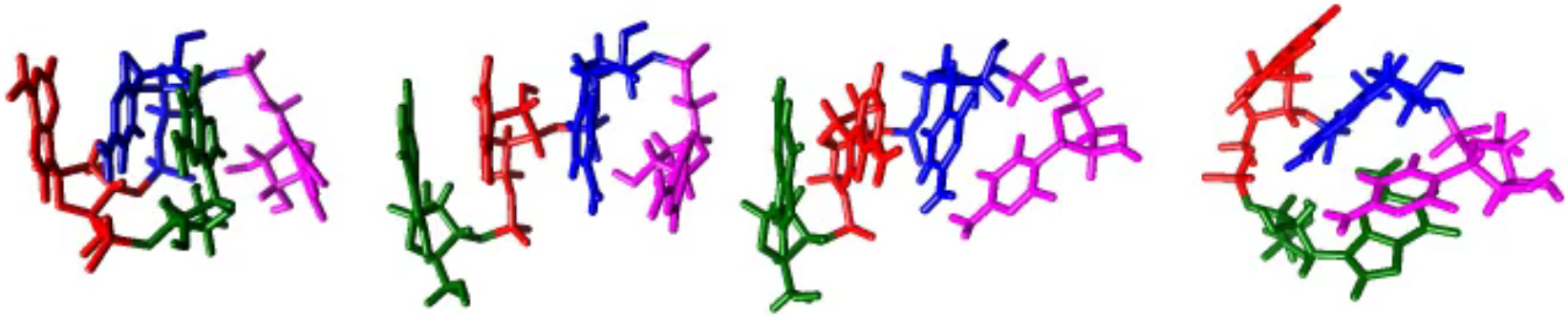
Today: two “long-time-to-develop” short stories…

✓ **can we converge DNA duplex structure/dynamics?**



Anonymous NIH R-01 reviewer in 2005:
“One has to wonder how many relatively short MD simulations have to be performed on short DNA fragments before what can be learned will have been learned…”

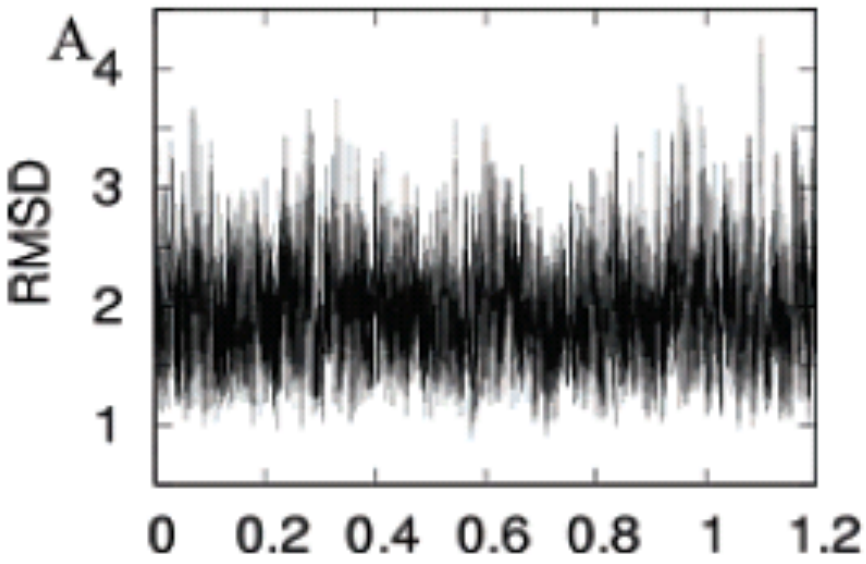
✓ **sampling RNA structure *accurately* is difficult**

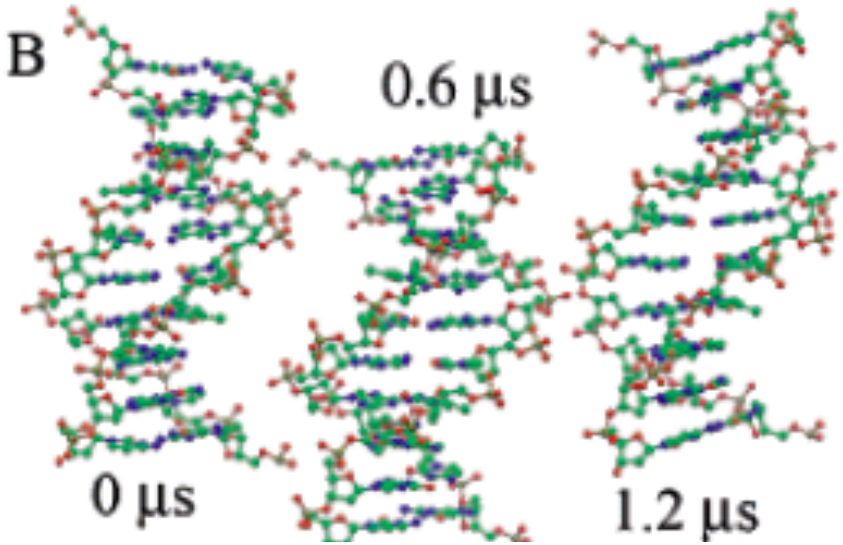


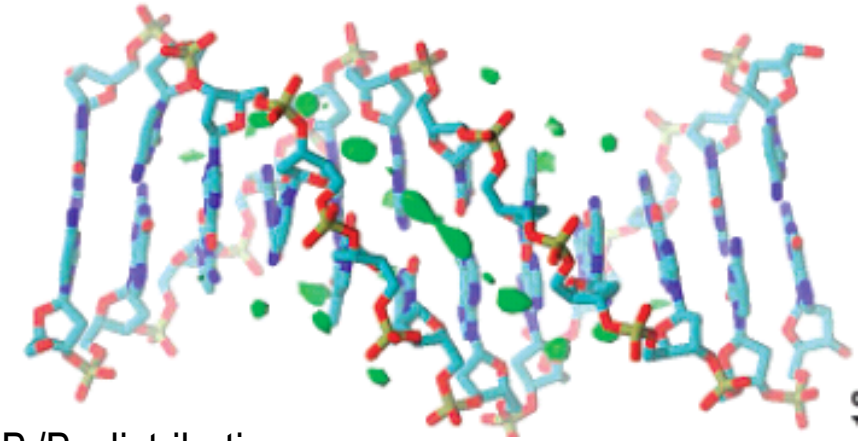
Dynamics of B-DNA on the Microsecond Time Scale

J. AM. CHEM. SOC. 2007, *129*, 14739−14745

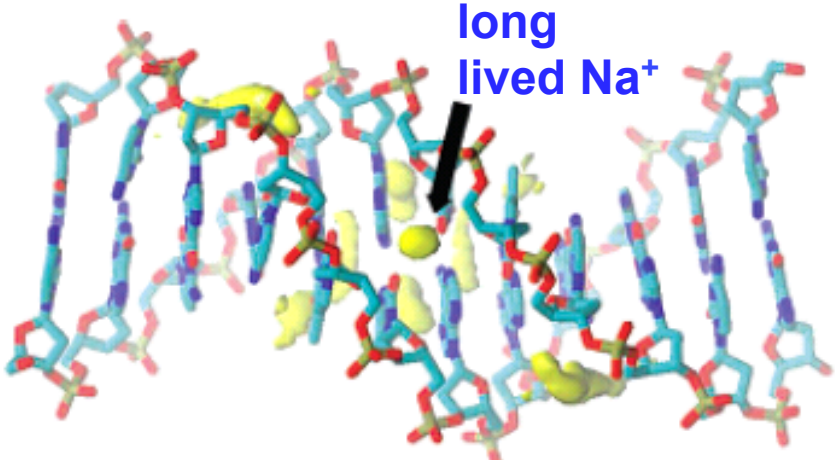
Alberto Pérez.[†,‡] F. Javier Luque.[§] and Modesto Orozco*[,†,‡,||]







B_I/B_{II} distributions still changing

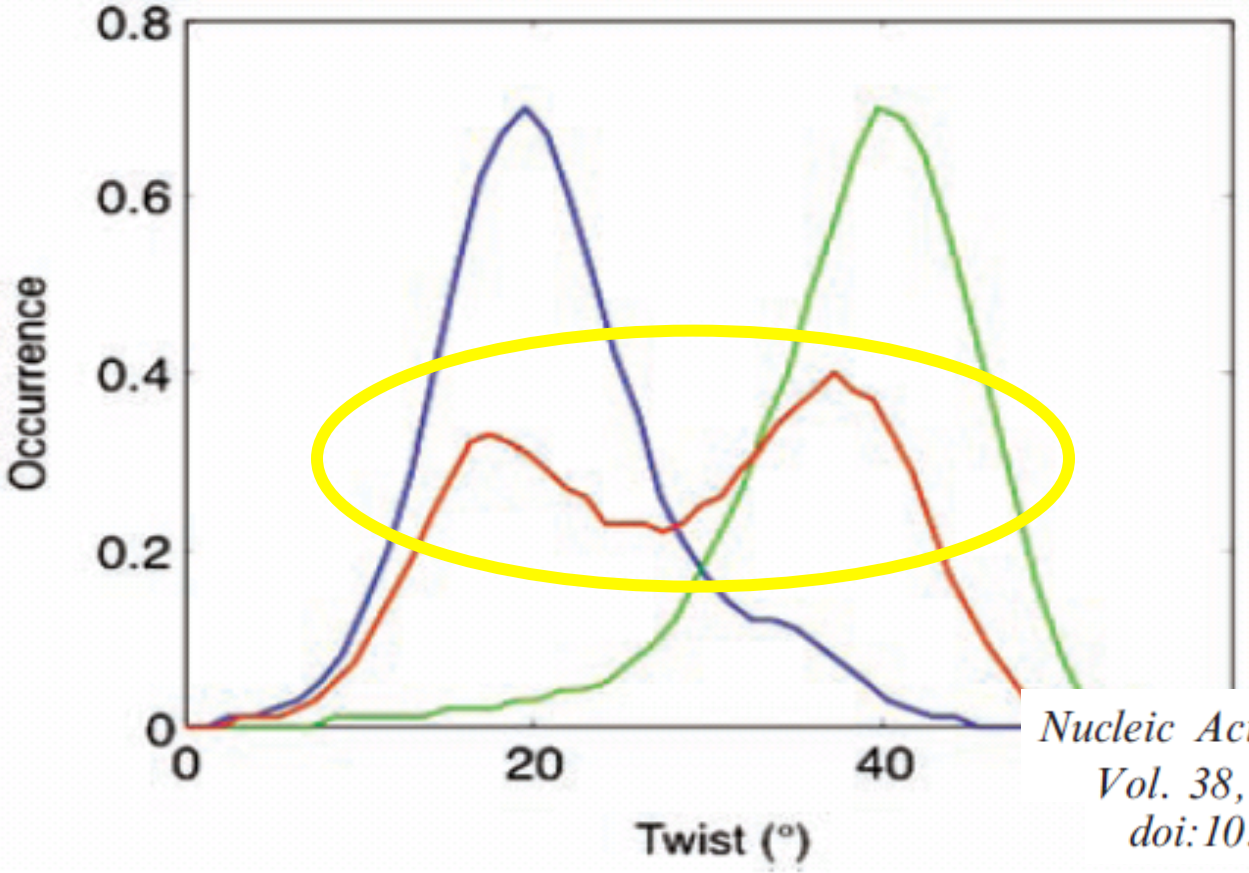


Convergence? Not yet…

Anton "testing" for ABC

ABC benchmark (50 ns, SPC/E + KCl)

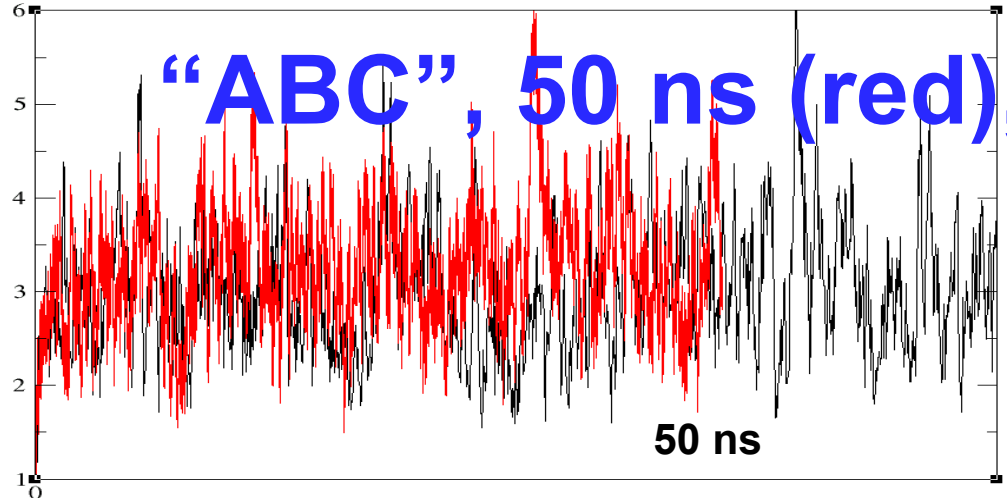
GAAC: GCACGAA**CG**AACGAACGC



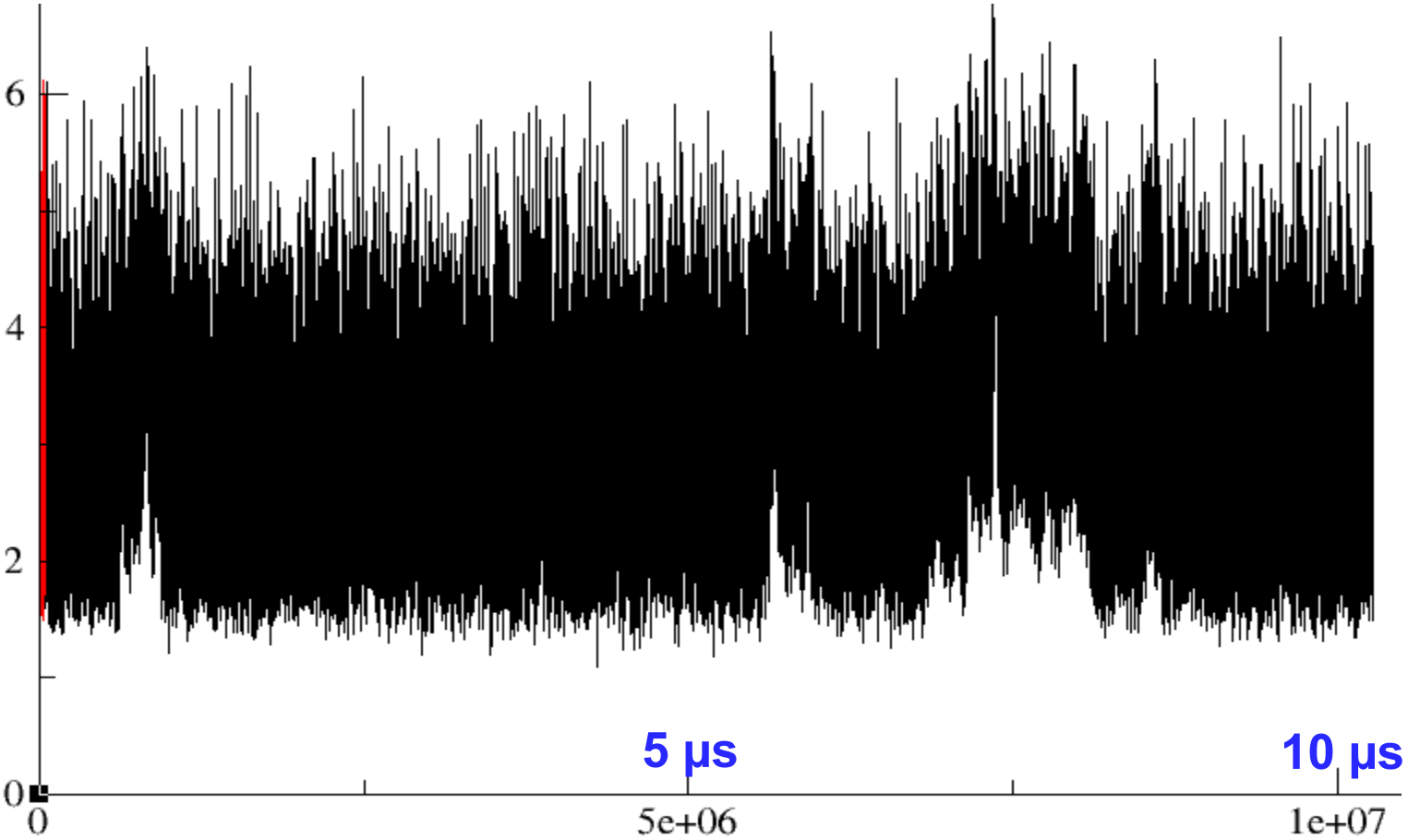
Nucleic Acids Research, 2010,
Vol. 38, No. 1 ***299–313***
doi:10.1093/nar/gkp834

Figure 6. Distribution of CG twist (degrees) as a function of the flanking sequences: CCGA (blue); ACGT (green); ACGA (red).

“ABC”, 50 ns (red), Anton (black)

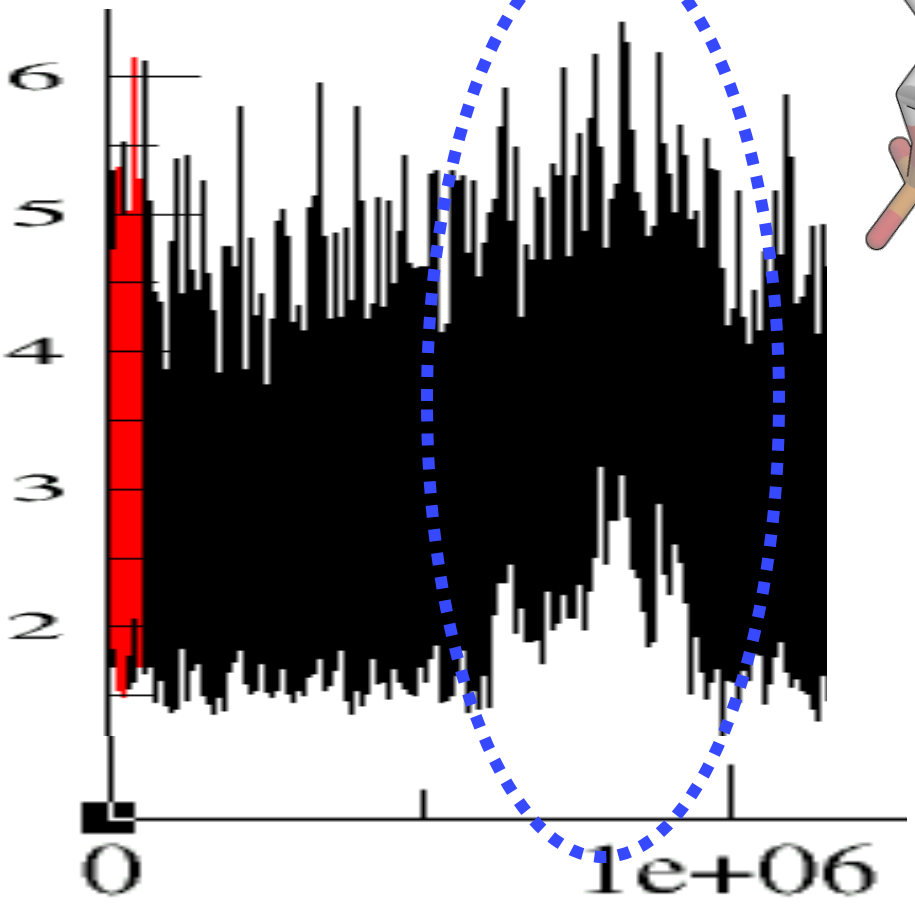


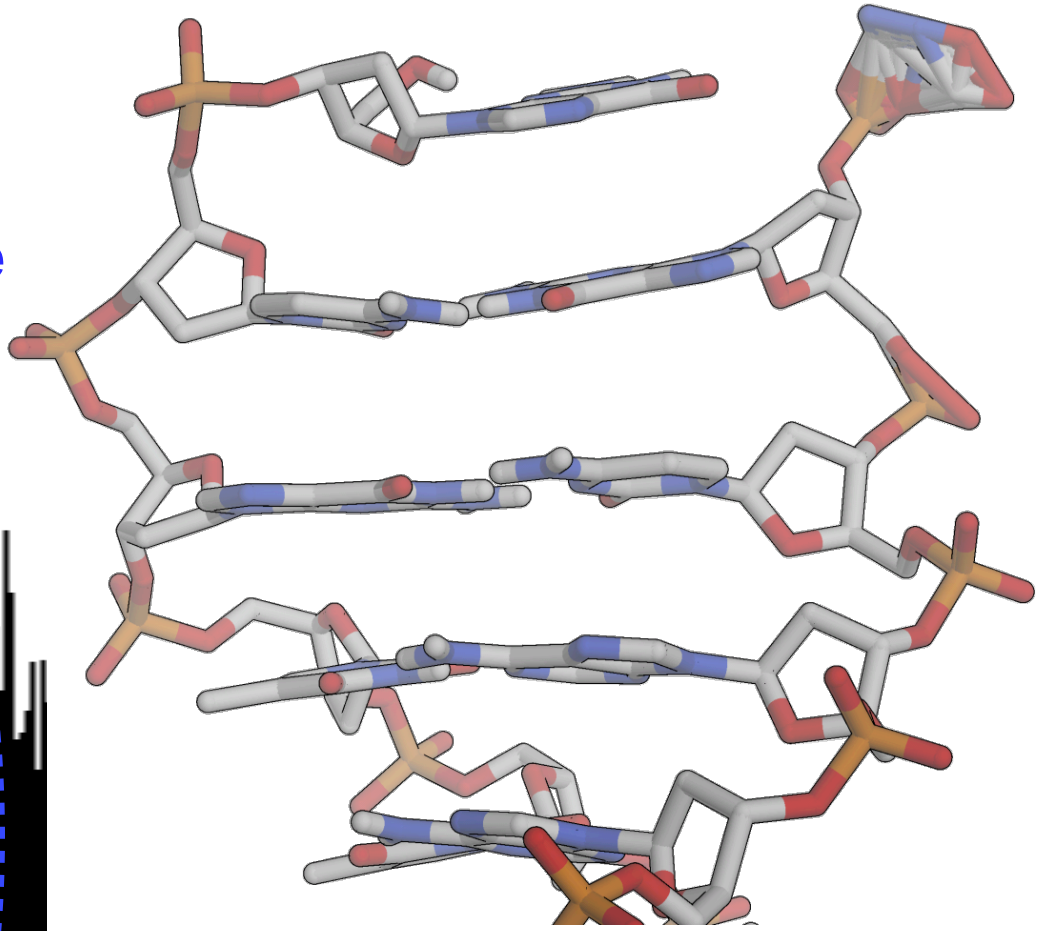
RMSD (Å) vs. time
Top: Original ABC work
Below: On Anton…



anton

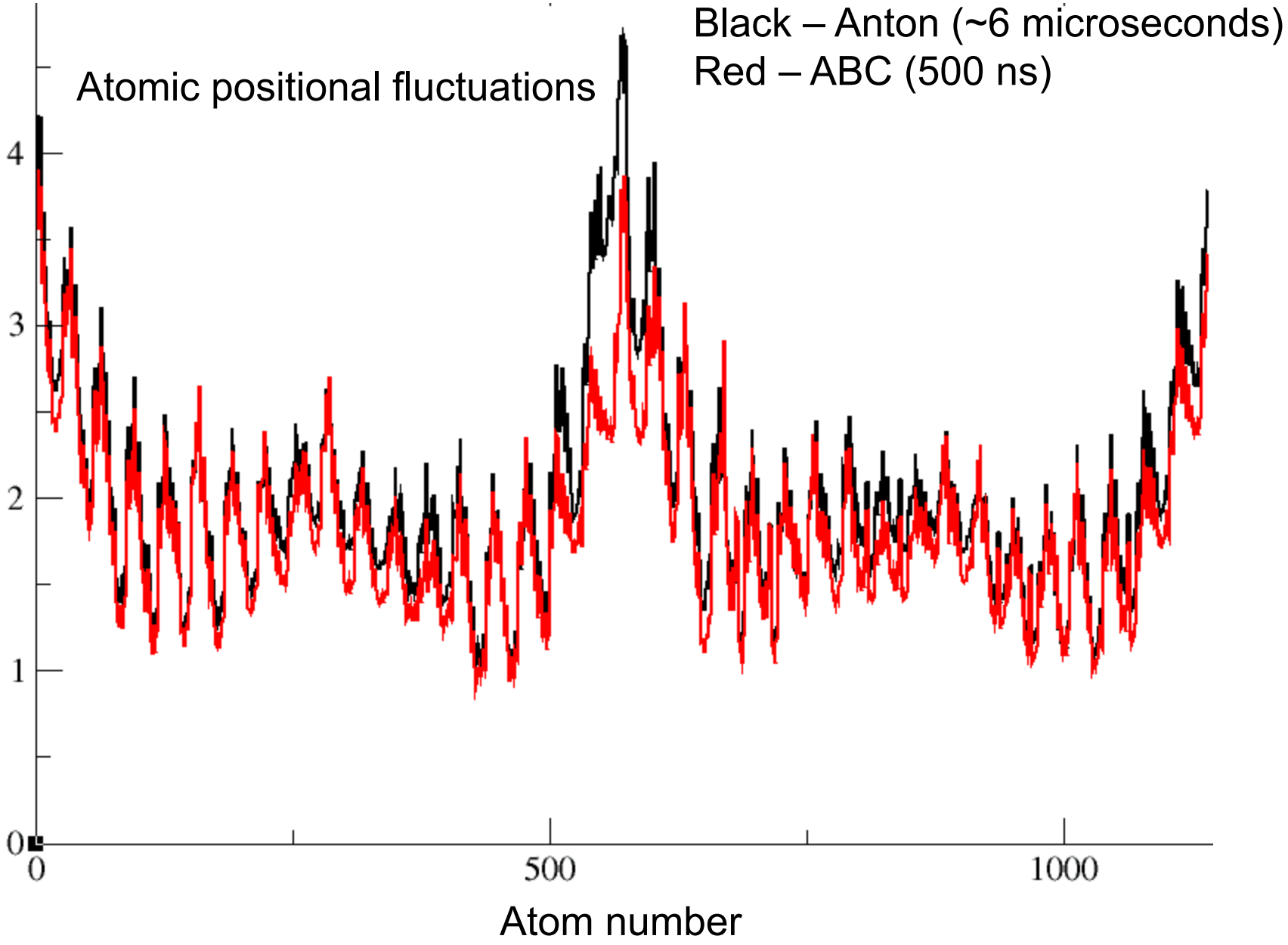
terminal base pair opening!



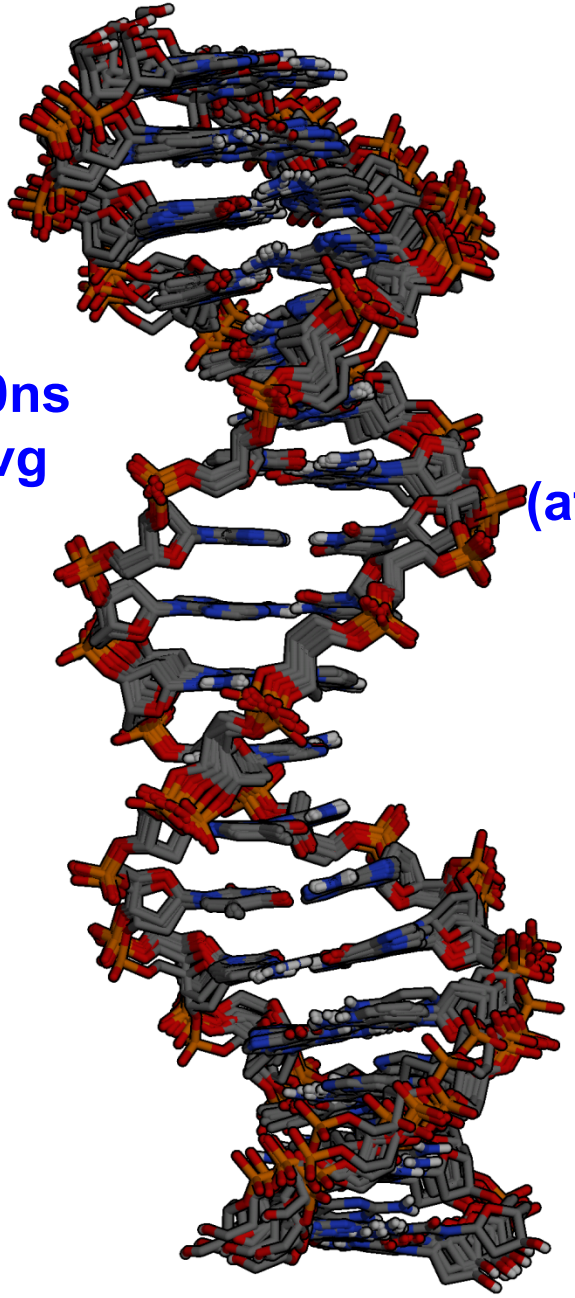


880-890 ns

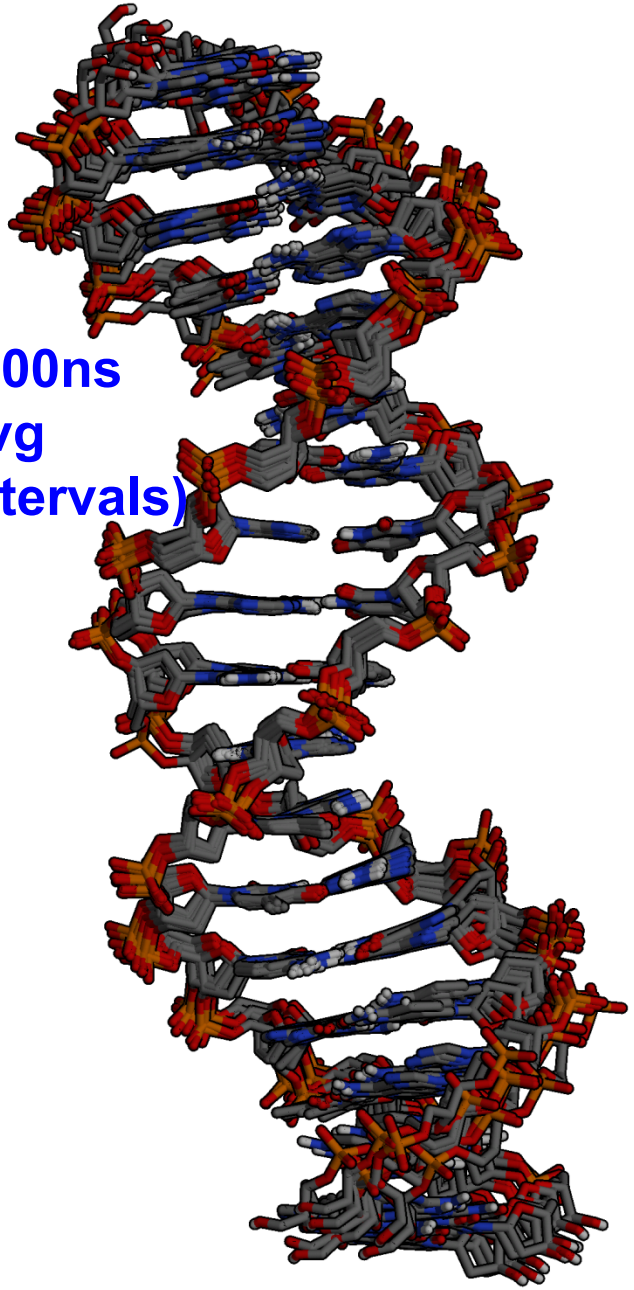
GAAC: GCACGAACGAACGAACGC



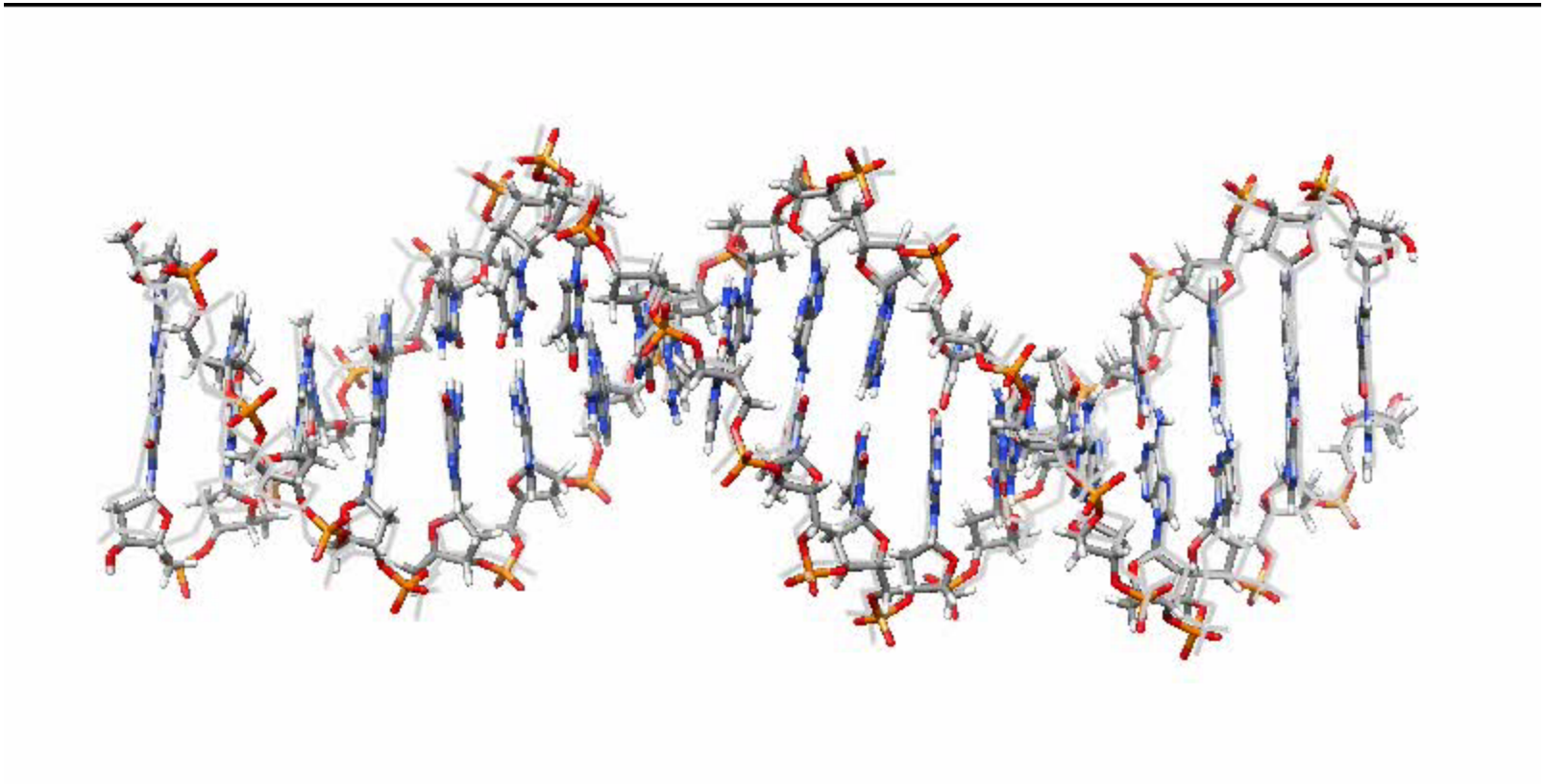
abc, 50ns
5ns avg



anton, 7000ns
5ns avg
(at 500ns intervals)



Anton run:



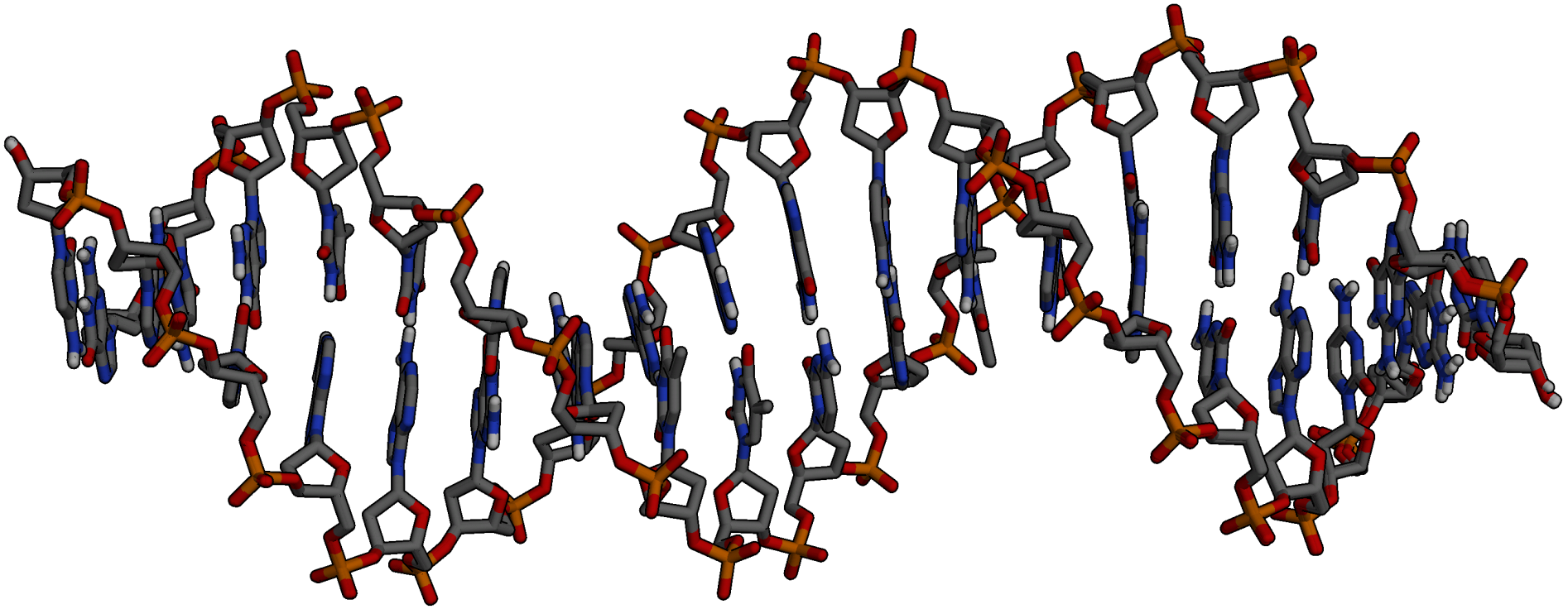
2 ns intervals, 10 ns running average, every 5th frame (~10 us).

~2010-2011

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)

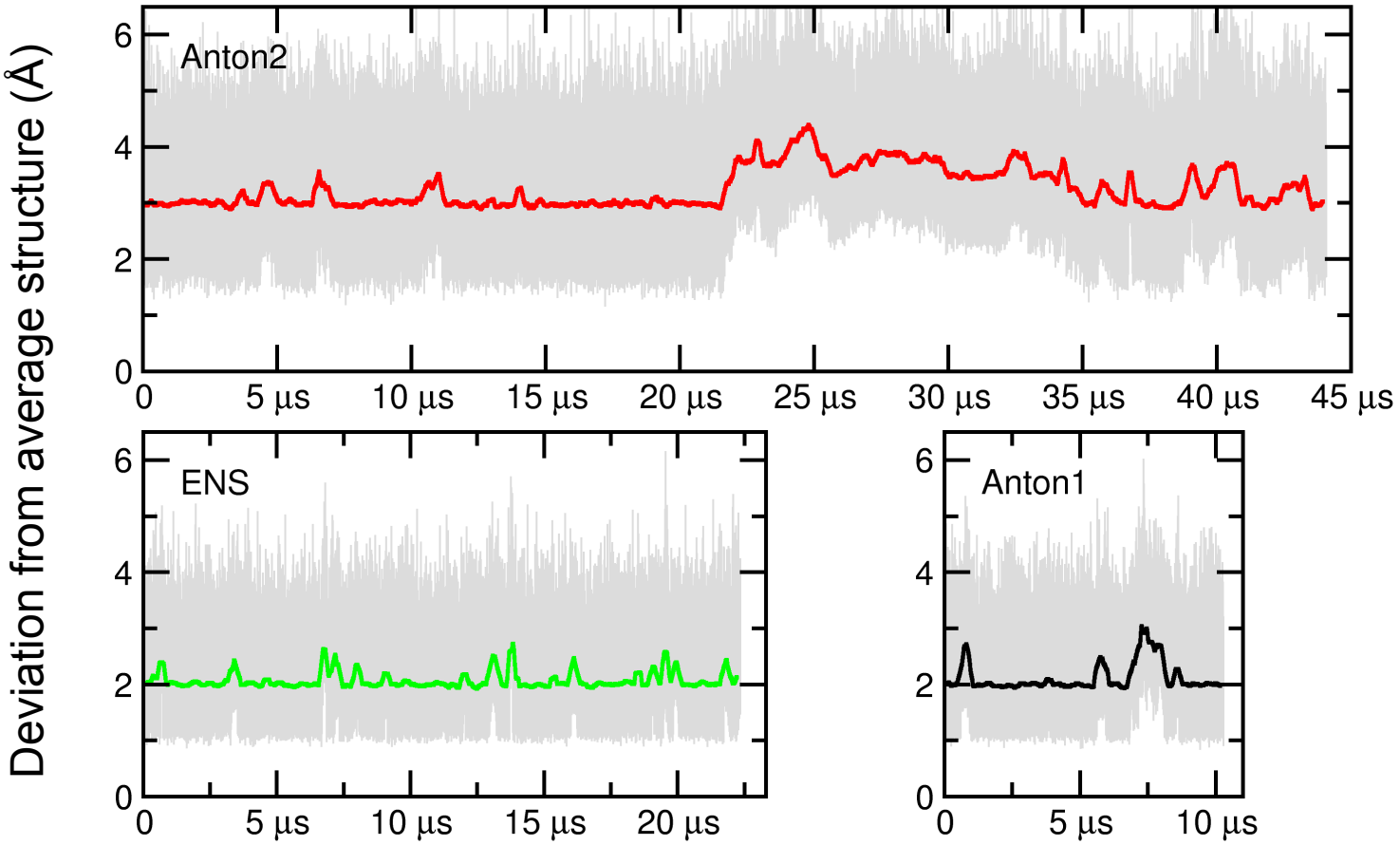


…this cannot be right, can it?

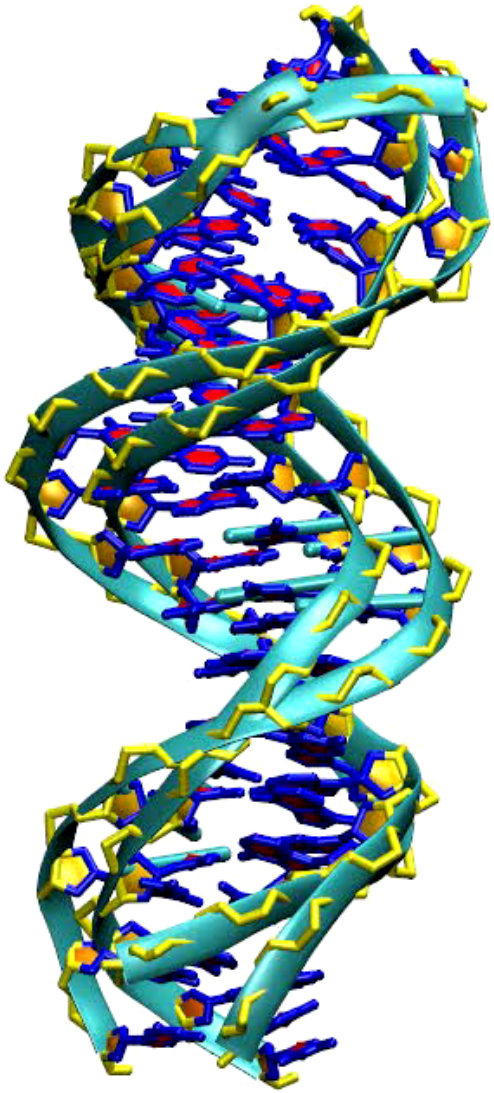
(breathing, bending, twisting, …)

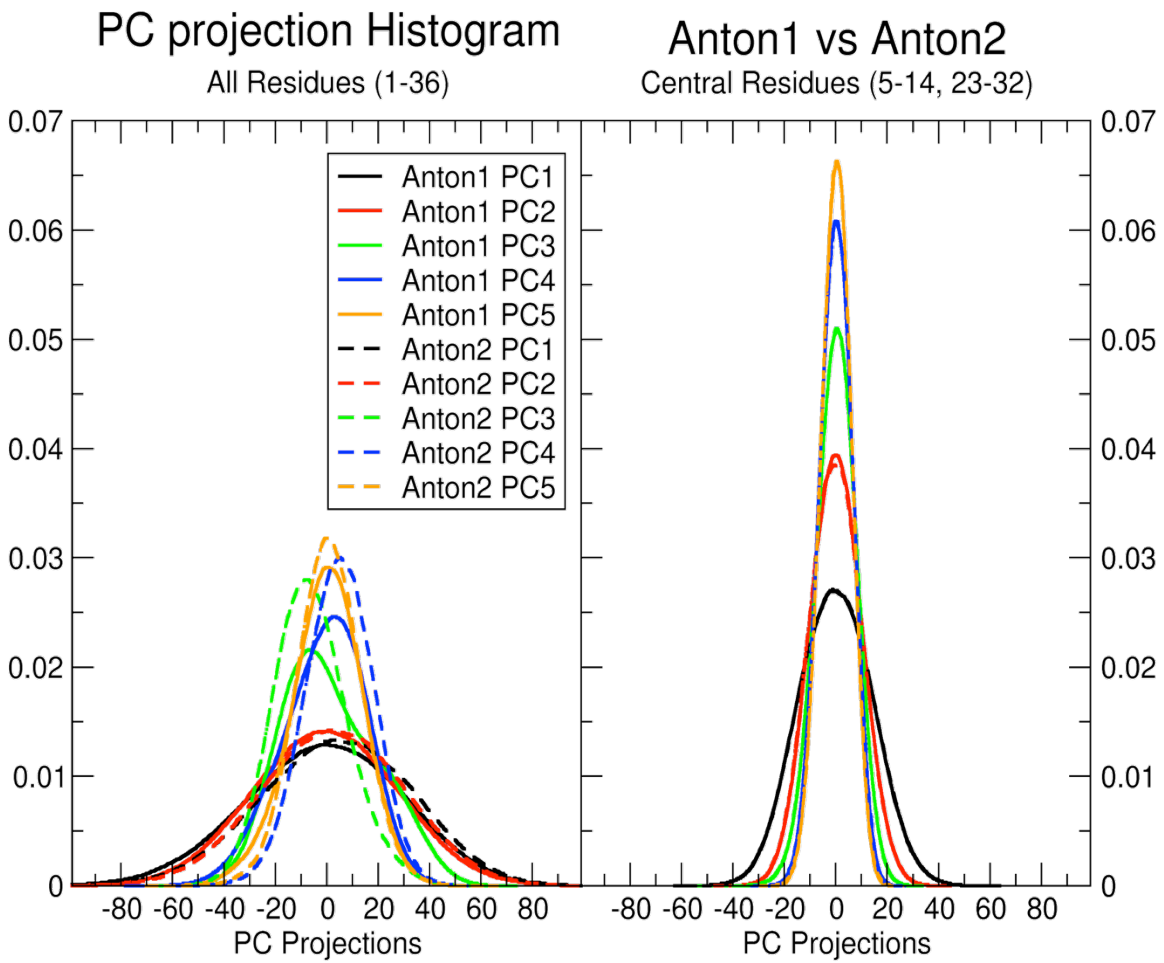
How to test?

- Do a longer run on Anton (write grant, get grant, run sims, ✔) = 44 μs
- Run an ensemble of 100 shorter simulations and aggregate = 20 μs
- Assume Anton is wrong: Run AMBER on CPUs and GPUs (~2 years, and still not long enough, only 2-4 μs ☹, but results are consistent ☺)

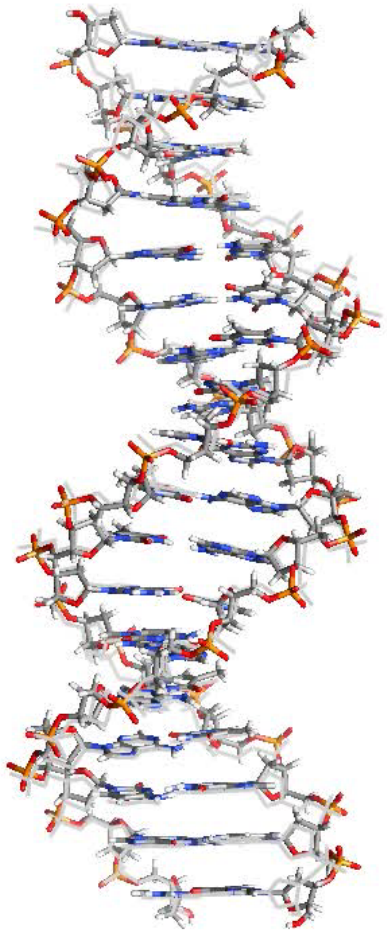


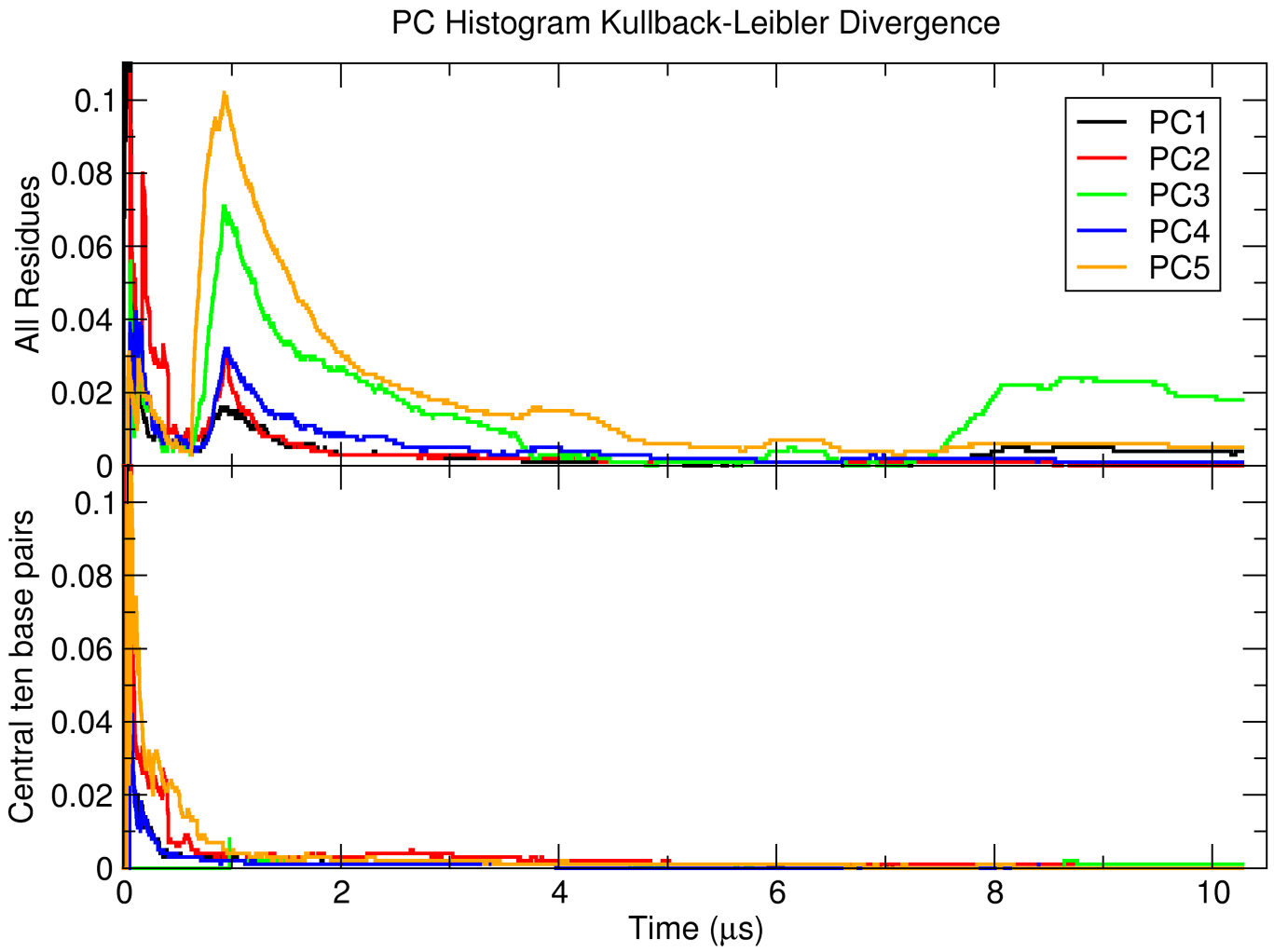
Test for convergence within and between simulations…



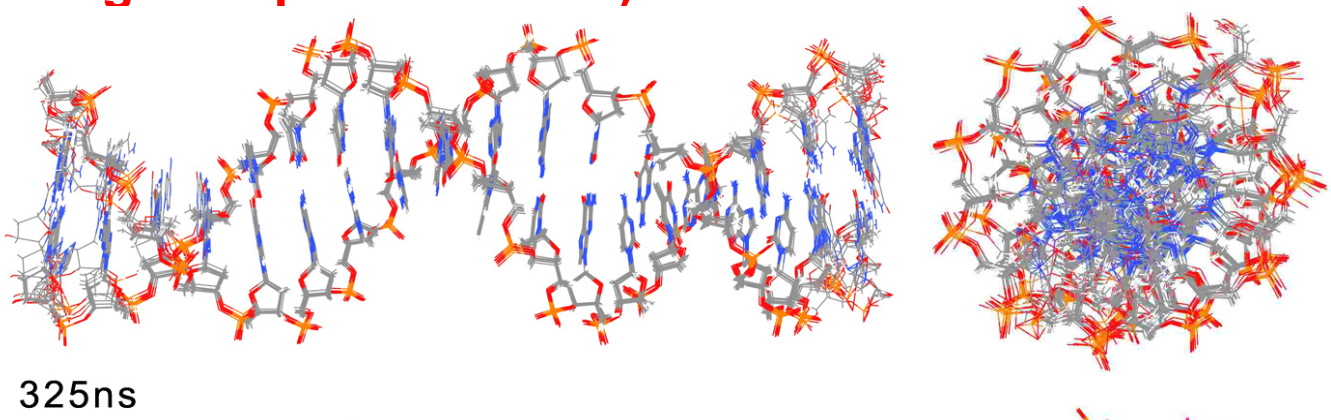


Test for convergence within and between simulations…

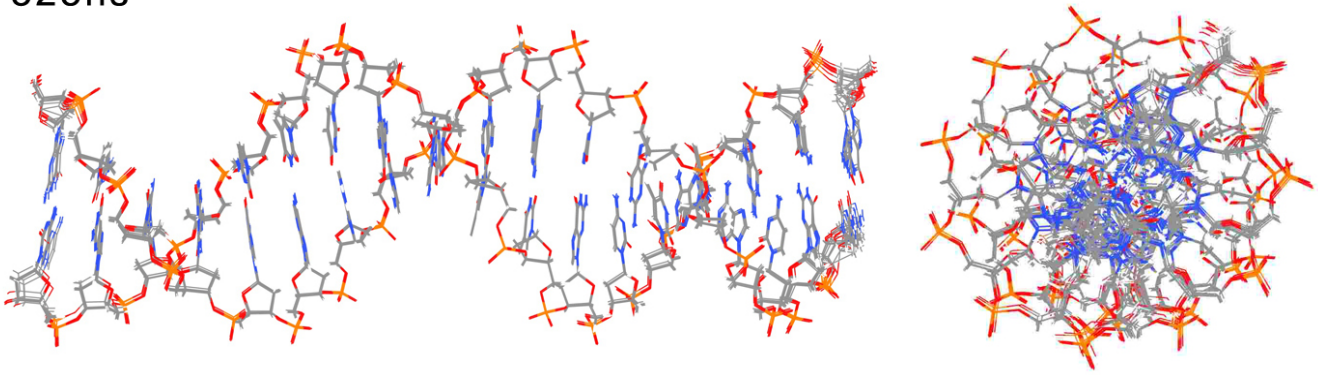




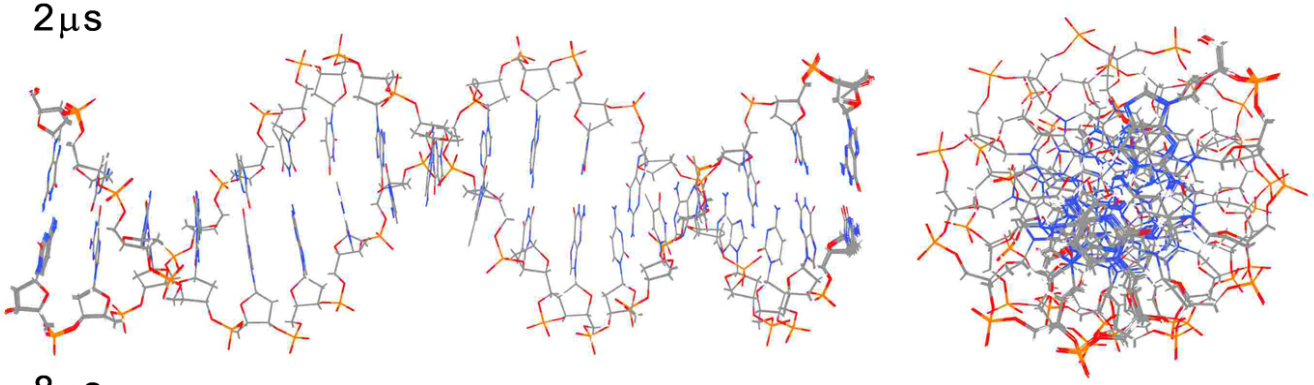
Test for convergence within and between simulations…
(perform running average over different timescales and cluster, showing 10 representatives)



325ns

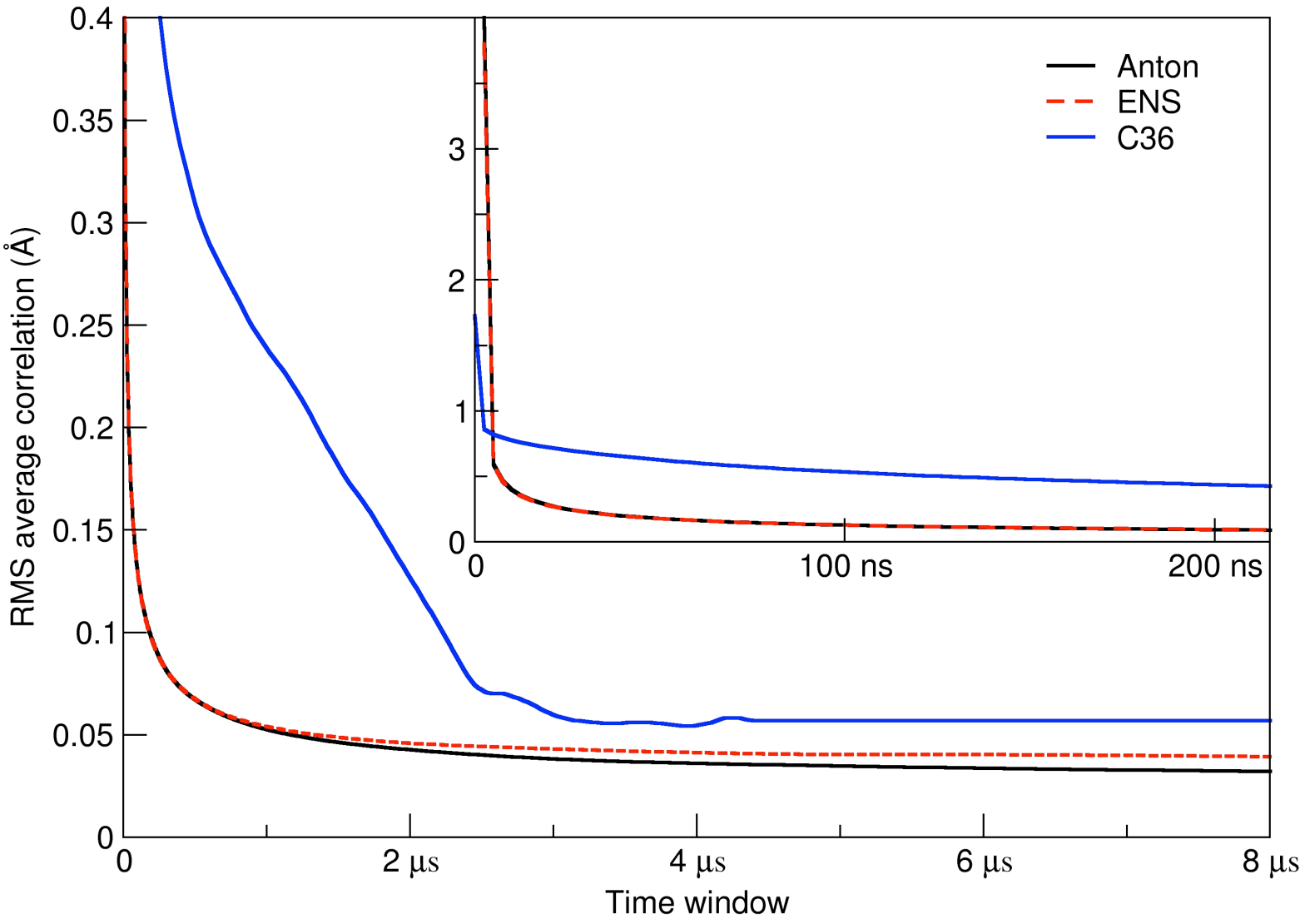


2μs

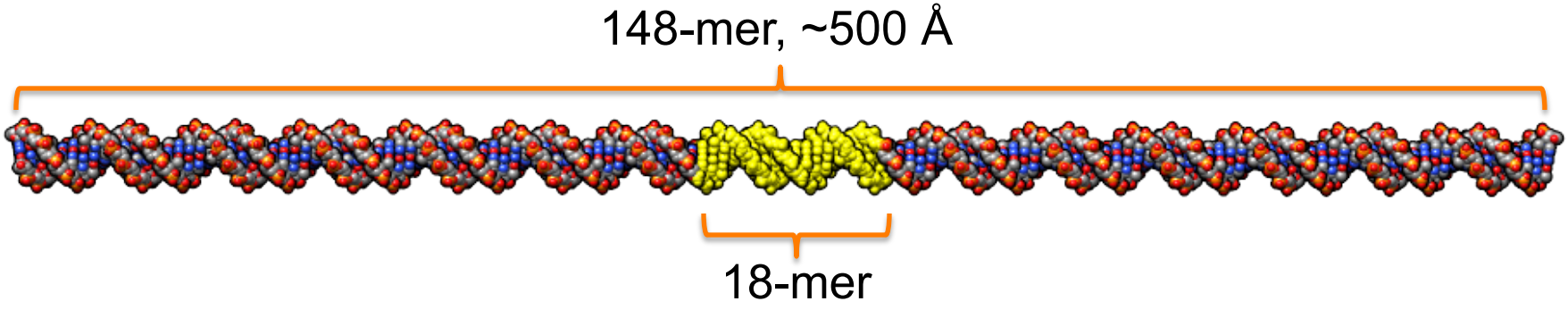


8μs

…alternative force field: CHARMM C36 runs on Blue Waters



DNA helices are relatively rigid, long persistence length

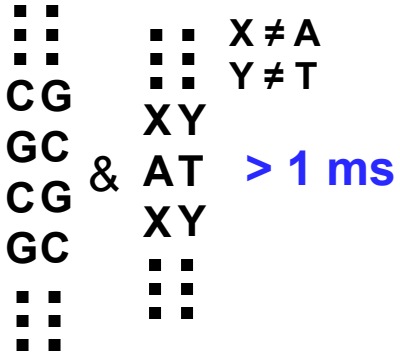


does it make sense for DNA to present consistent structure and for regular Watson-Crick DNA to be “rigid” – YES!

Questions about recognition:

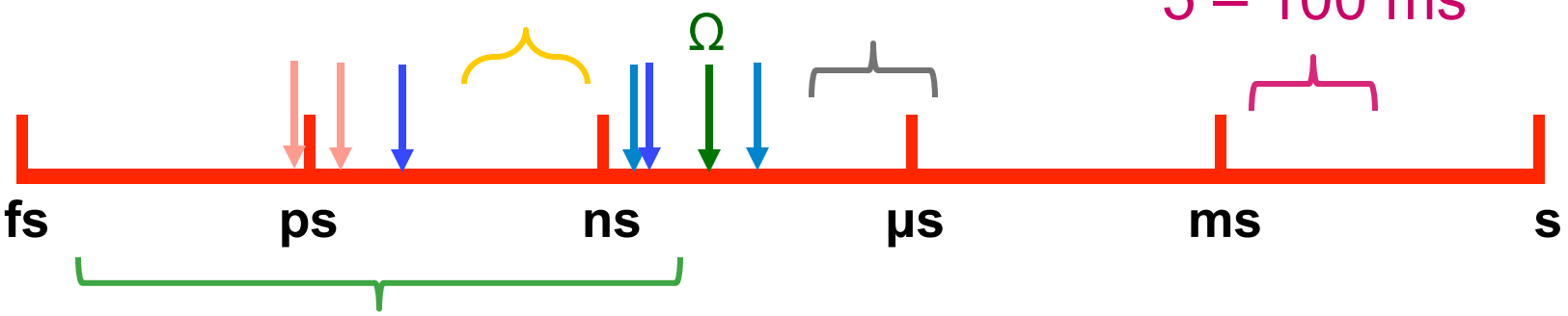
- conformational selection?
- induced fit / deformability?
- why are mismatches easily recognized?

What do we know about the dynamics of DNA helices?



^{13}C & ^{15}N NMR
triplet anisotropy decay
electron paramagnetic resonance
PELDOR
31P and/or field-cycling NMR
FT IR difference spectroscopy

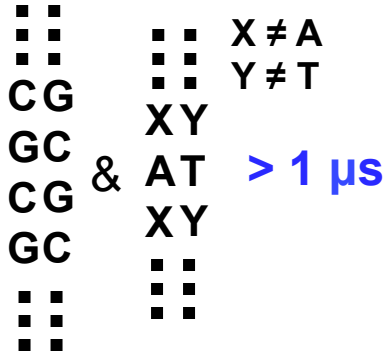
NMR: base pair opening
5 – 100 ms



Berg: Dynamic stokes shift (base replaced by dye)
"power law" dynamics over 6 orders of magnitude of time
40 fs – 40 ns

What about longer timescales?

What do we know about the dynamics of DNA helices?



^{13}C & ^{15}N NMR

triplet anisotropy decay

electron paramagnetic resonance

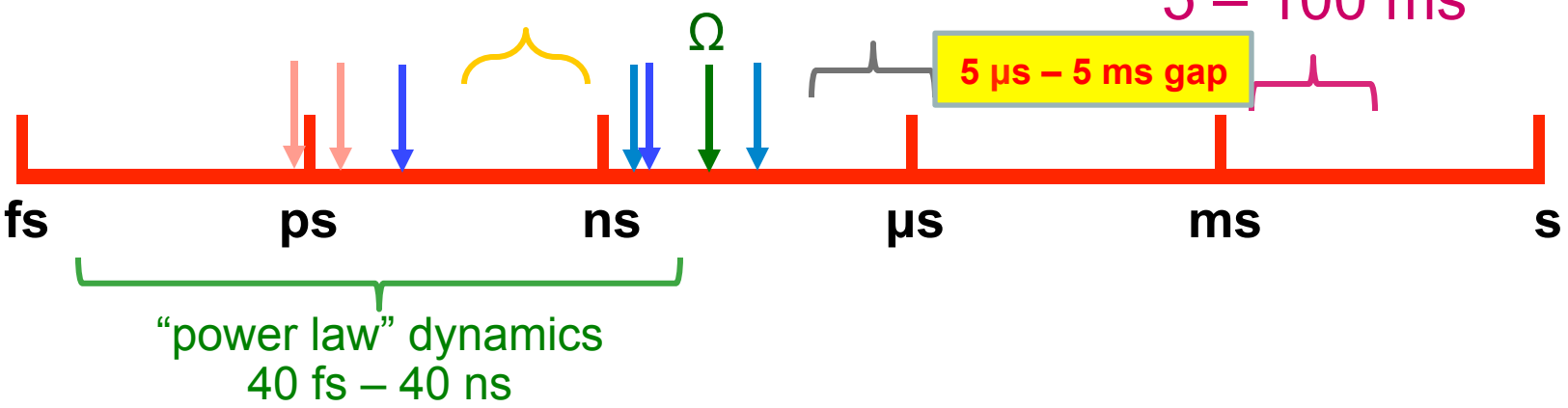
PELDOR

31P and/or field-cycling NMR

FT IR difference spectroscopy

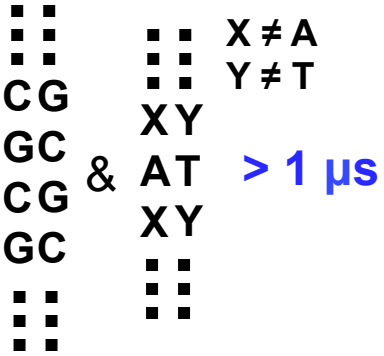
NMR: base pair opening

5 – 100 ms



is this “gap” in dynamics real?

What do we know about the dynamics of DNA helices?



^{13}C & ^{15}N NMR

triplet anisotropy decay

electron paramagnetic resonance

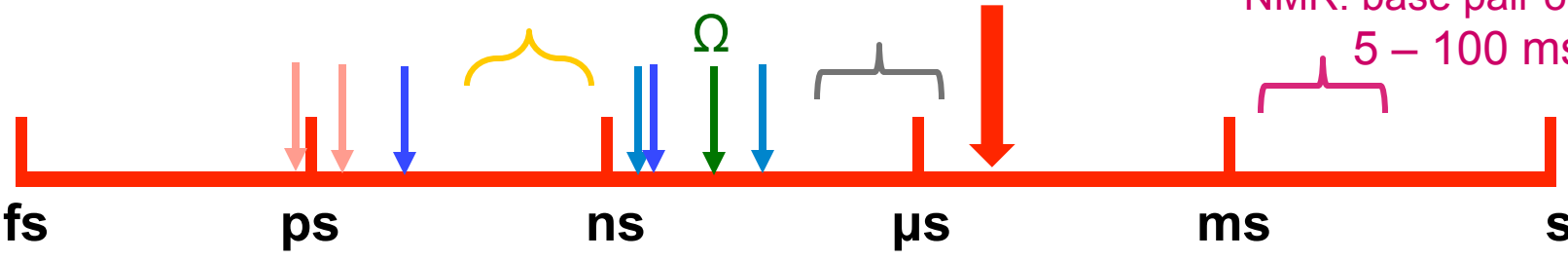
PELDOR

31P and/or field-cycling NMR

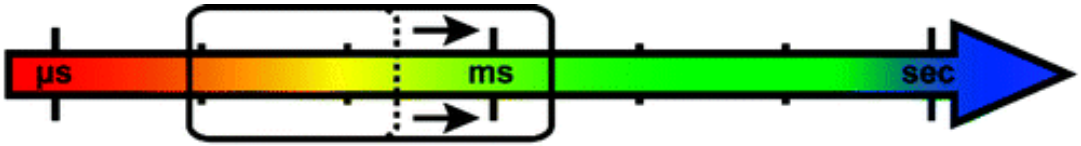
FT IR difference spectroscopy

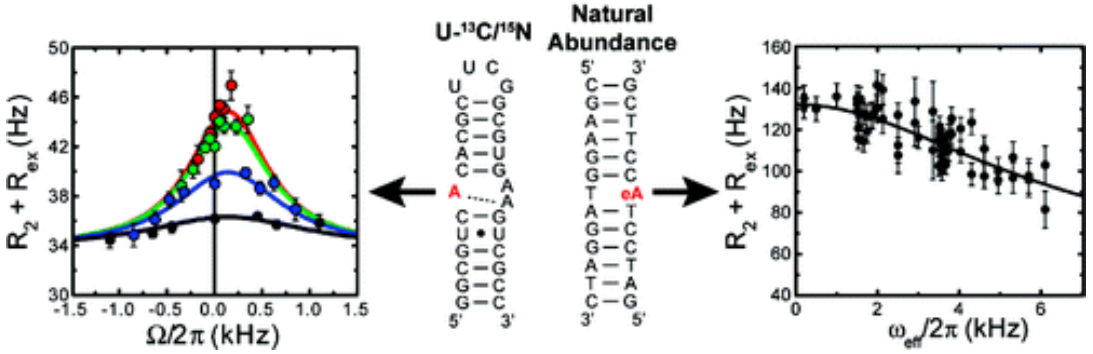
selective off-resonance
$R_{1\rho}$ carbon relaxation
26 ± 8 µs
for mismatches!

NMR: base pair opening
5 – 100 ms

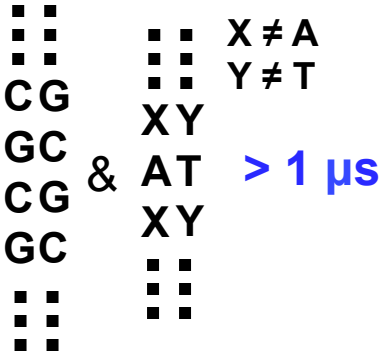


"power law" dynamics
40 fs – 40 ns





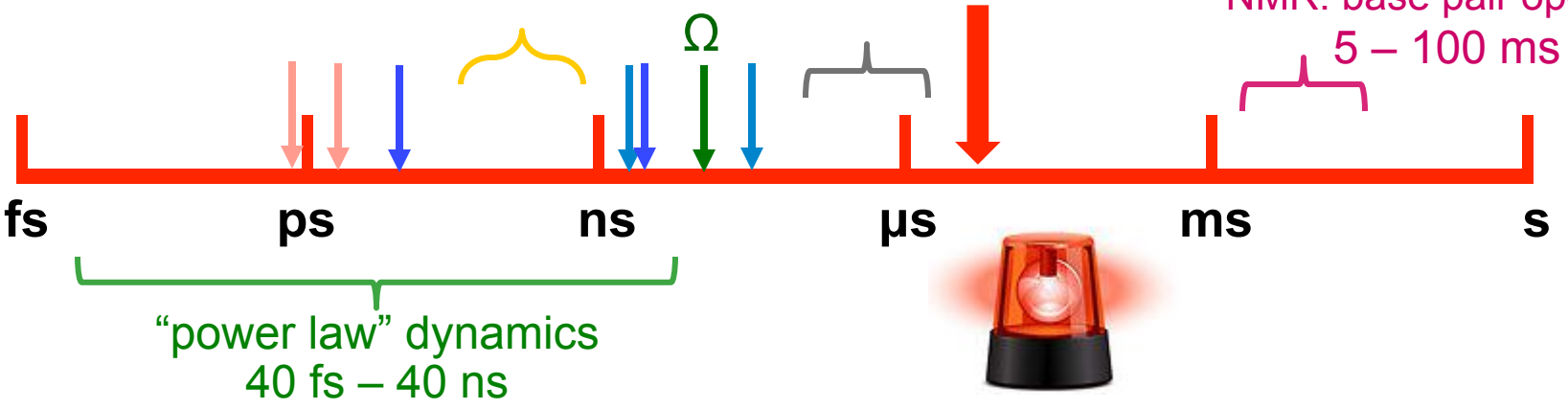
What do we know about the dynamics of DNA helices?



^{13}C & ^{15}N NMR
triplet anisotropy decay
electron paramagnetic resonance
PELDOR
31P and/or field-cycling NMR
FT IR difference spectroscopy

selective off-resonance
$R_{1\rho}$ carbon relaxation
26 ± 8 µs
for mismatches!

NMR: base pair opening
5 – 100 ms

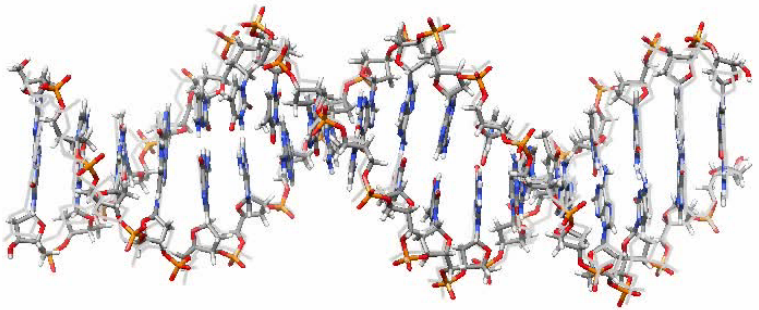


Questions about recognition:

- conformational selection? — no, decay is too fast!
- induced fit / deformability? — requires bp opening
- why are mismatches easily recognized? — timescale mismatch

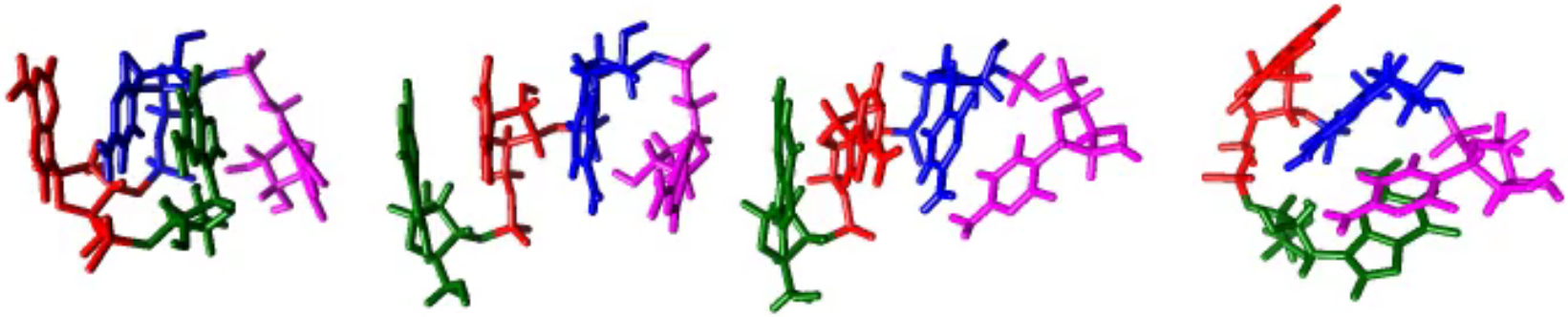
Today: two “long-time-to-develop” short stories…

✓ **can we converge DNA duplex structure/dynamics?**



Anonymous NIH R-01 reviewer in 2005:
“One has to wonder how many relatively short MD simulations have to be performed on short DNA fragments before what can be learned will have been learned…”

✓ **sampling RNA structure *accurately* is difficult**



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

all tetraloops

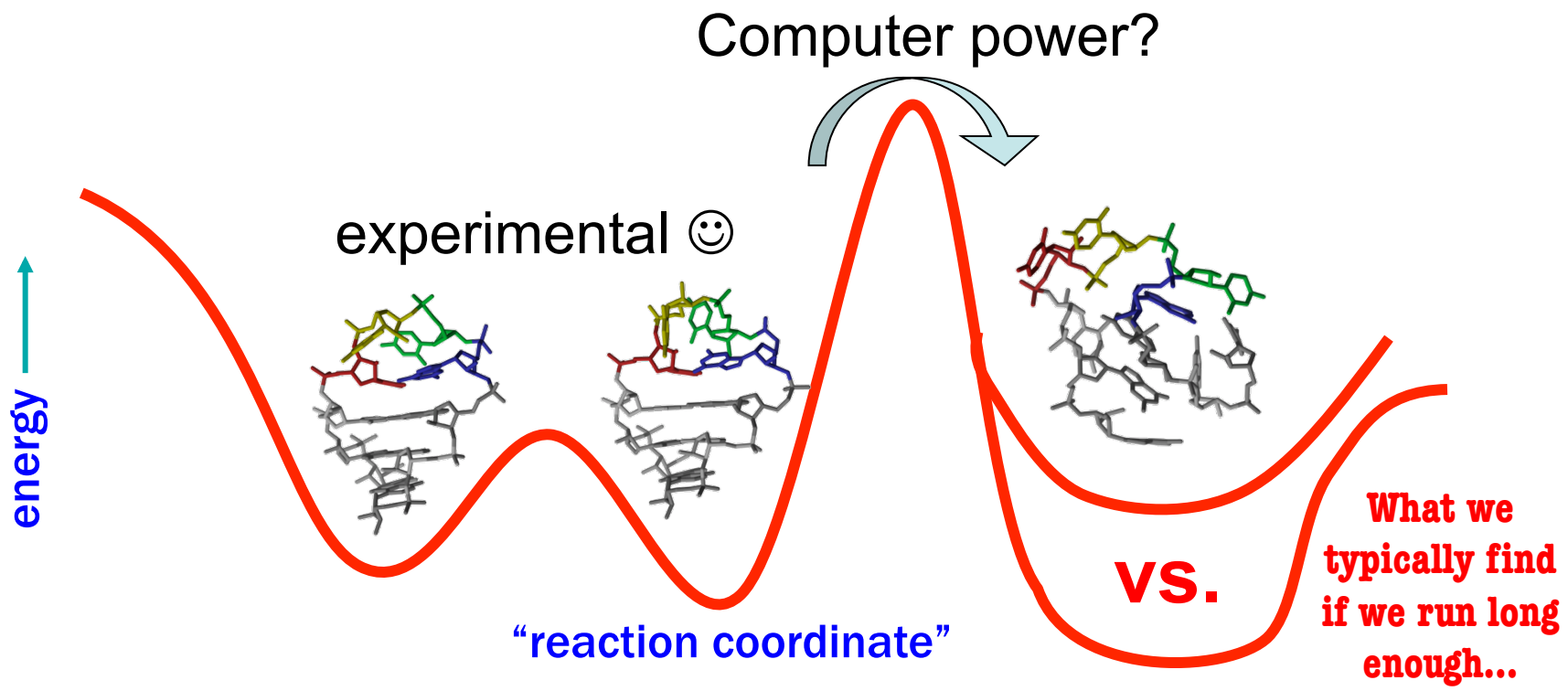
NMR structures of DNA & RNA

crystal simulations

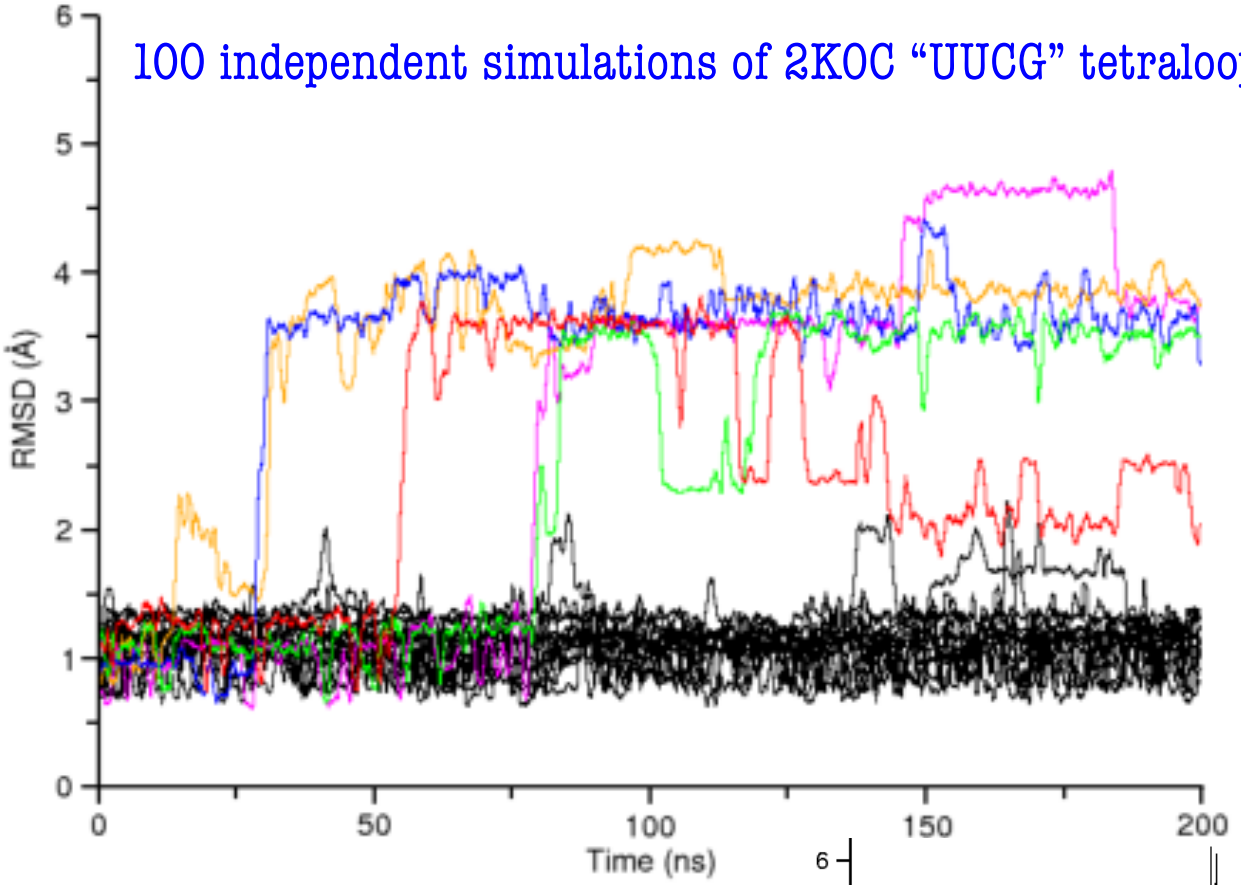
RNA motifs

quadruplexes

RNA-drug interactions

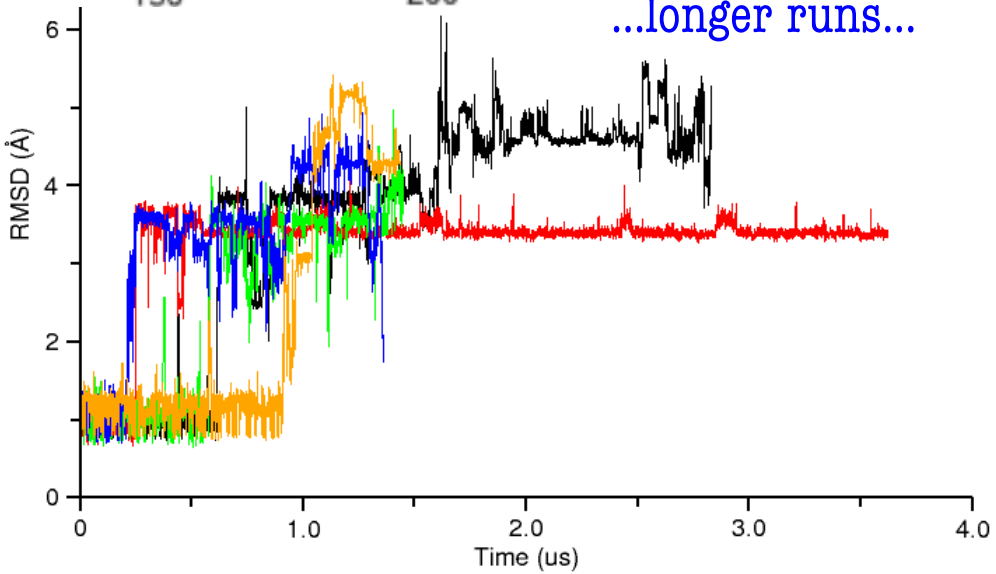


100 independent simulations of 2KOC “UUCG” tetraloop



Limited sampling & too complex: Is there a simpler set of systems?

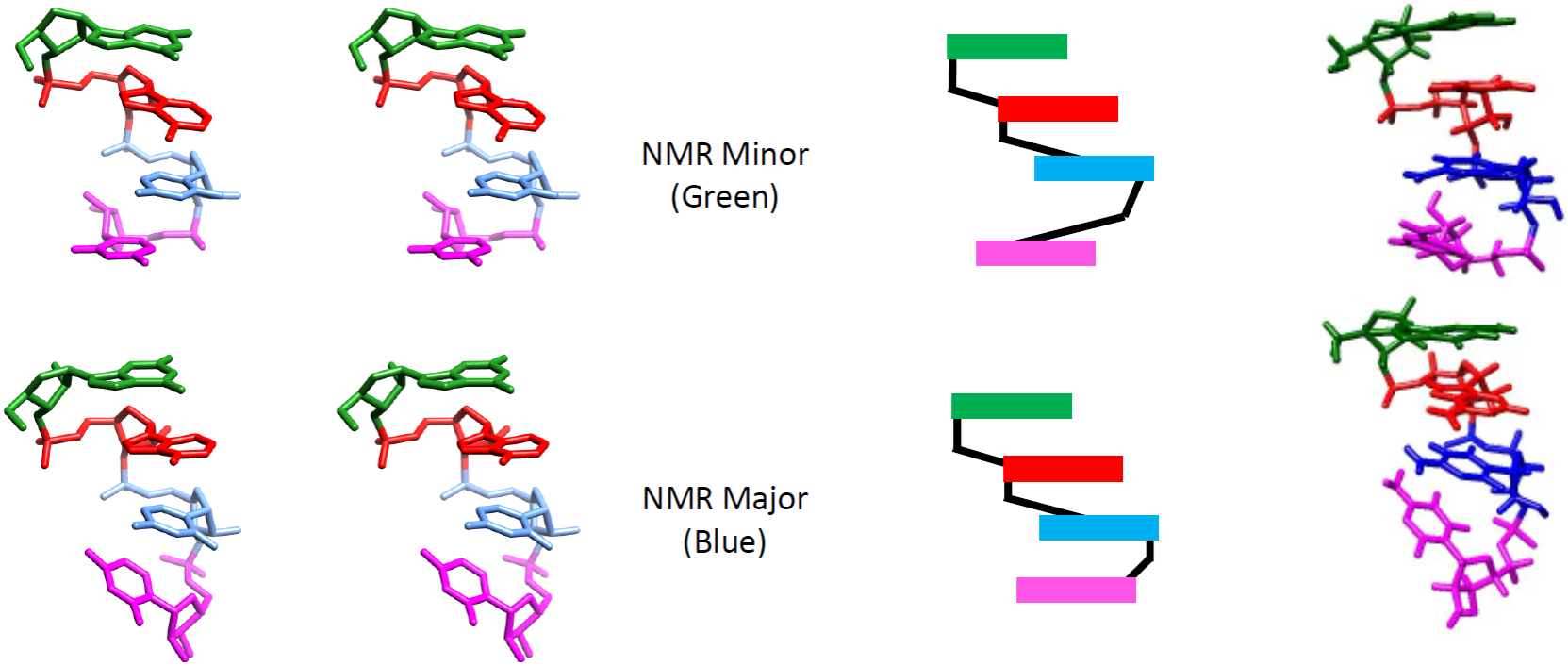
...longer runs...



…a system where we can get complete sampling

r(GACC) tetranucleotide

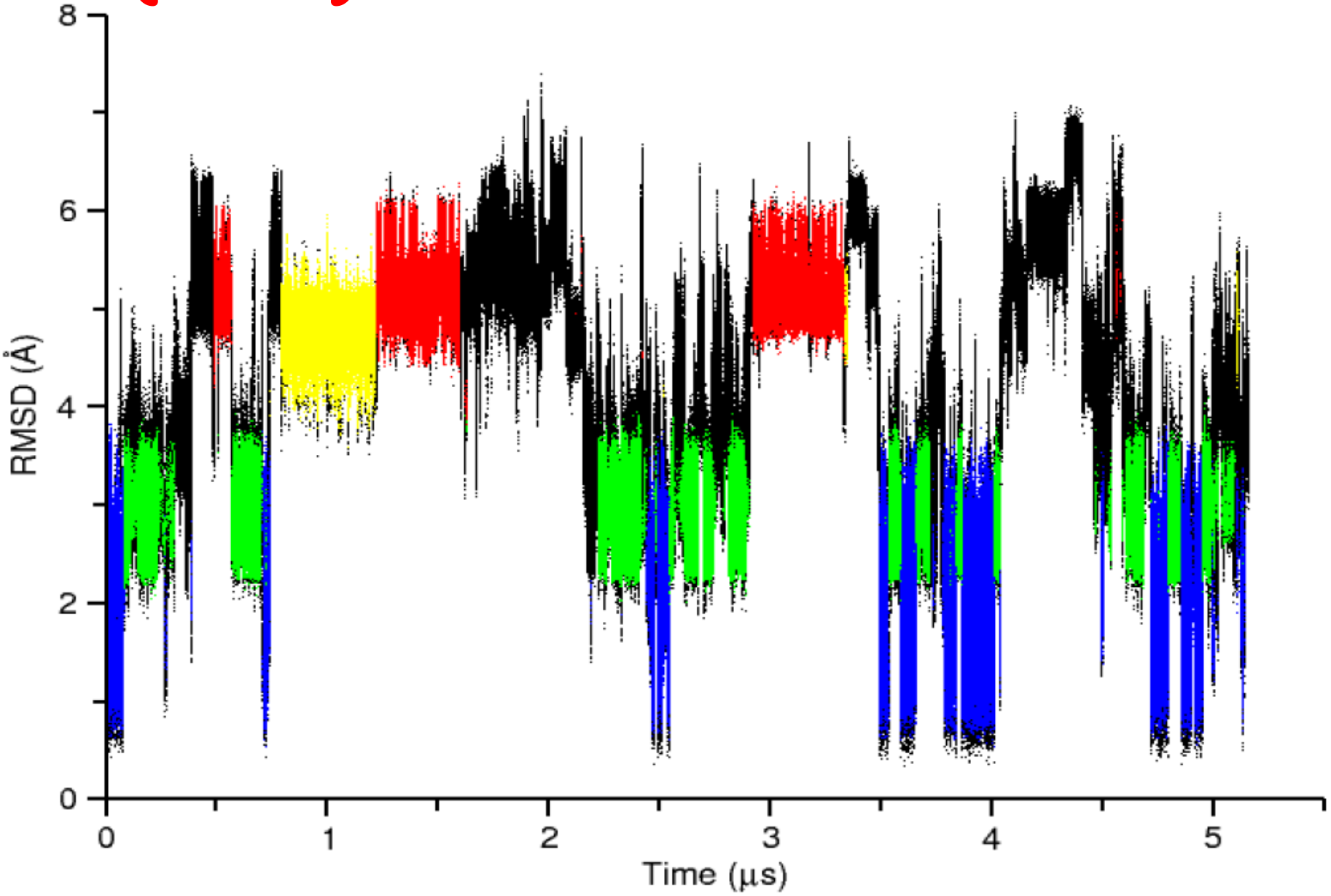
[Turner / Yildirim]



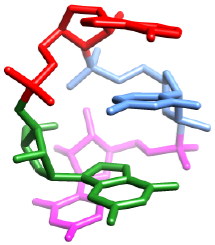
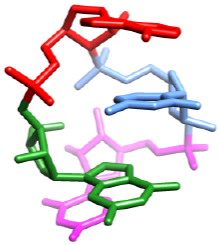
NMR suggests two dominant conformations…

…compare to MD simulations in explicit solvent

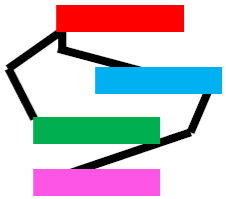
r(GACC) tetranucleotide: AMBER ff12



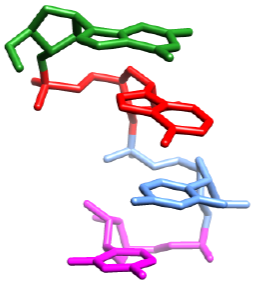
< explicit solvent time-contiguous MD >

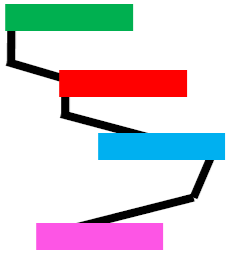
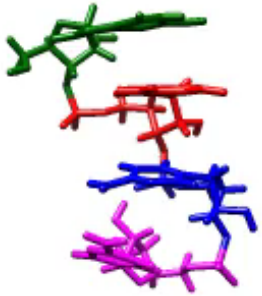
Intercalated
(Red)

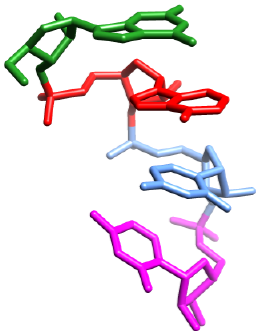
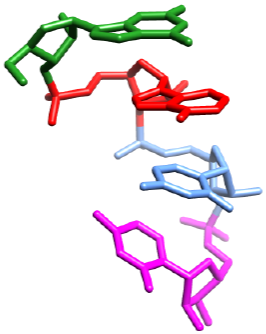




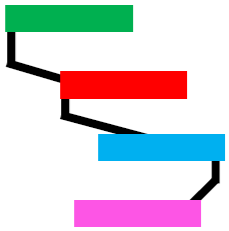
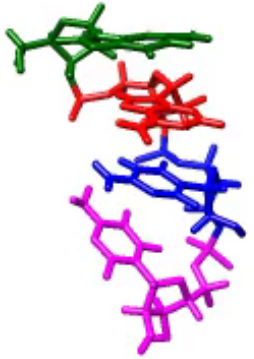


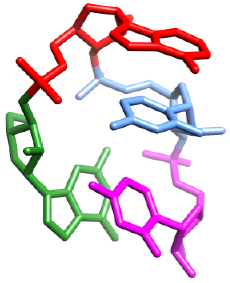
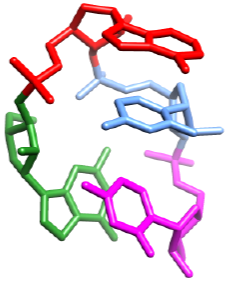
NMR Minor
(Green)

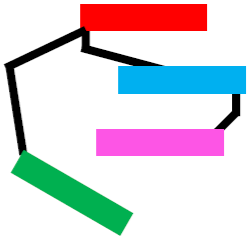
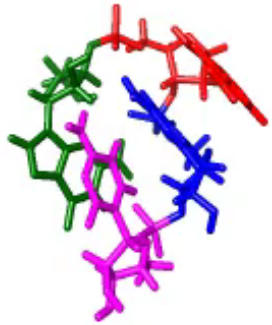



NMR Major
(Blue)

Inverted
(Yellow)

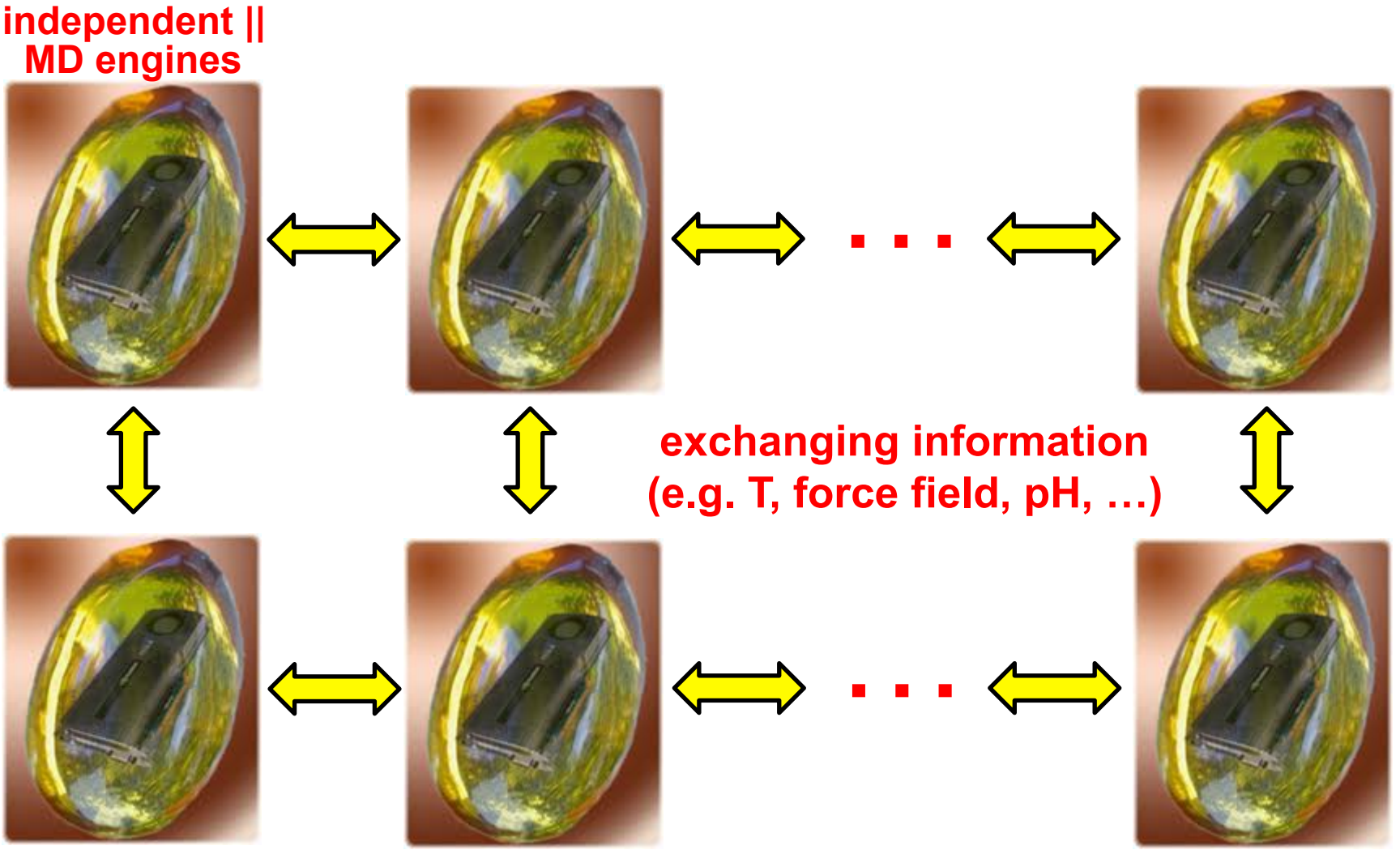



...still need more sampling!

(enablers)

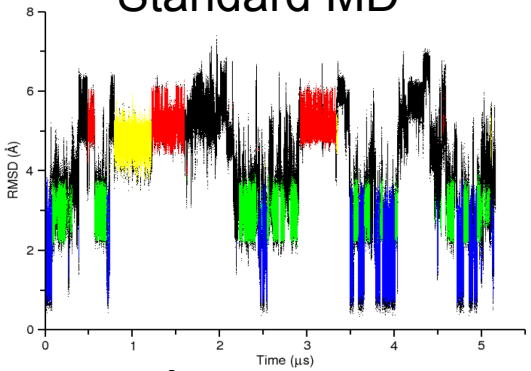
- strong GPU performance of AMBER/PMEMD
- good replica exchange functionality
- access to Keeneland, Stampede, Blue Waters, ...

Blue Waters PRAC: The main goals are to hierarchically and tightly couple a series of optimized molecular dynamics engines to fully map out the conformational, energetic and chemical landscape of RNA.



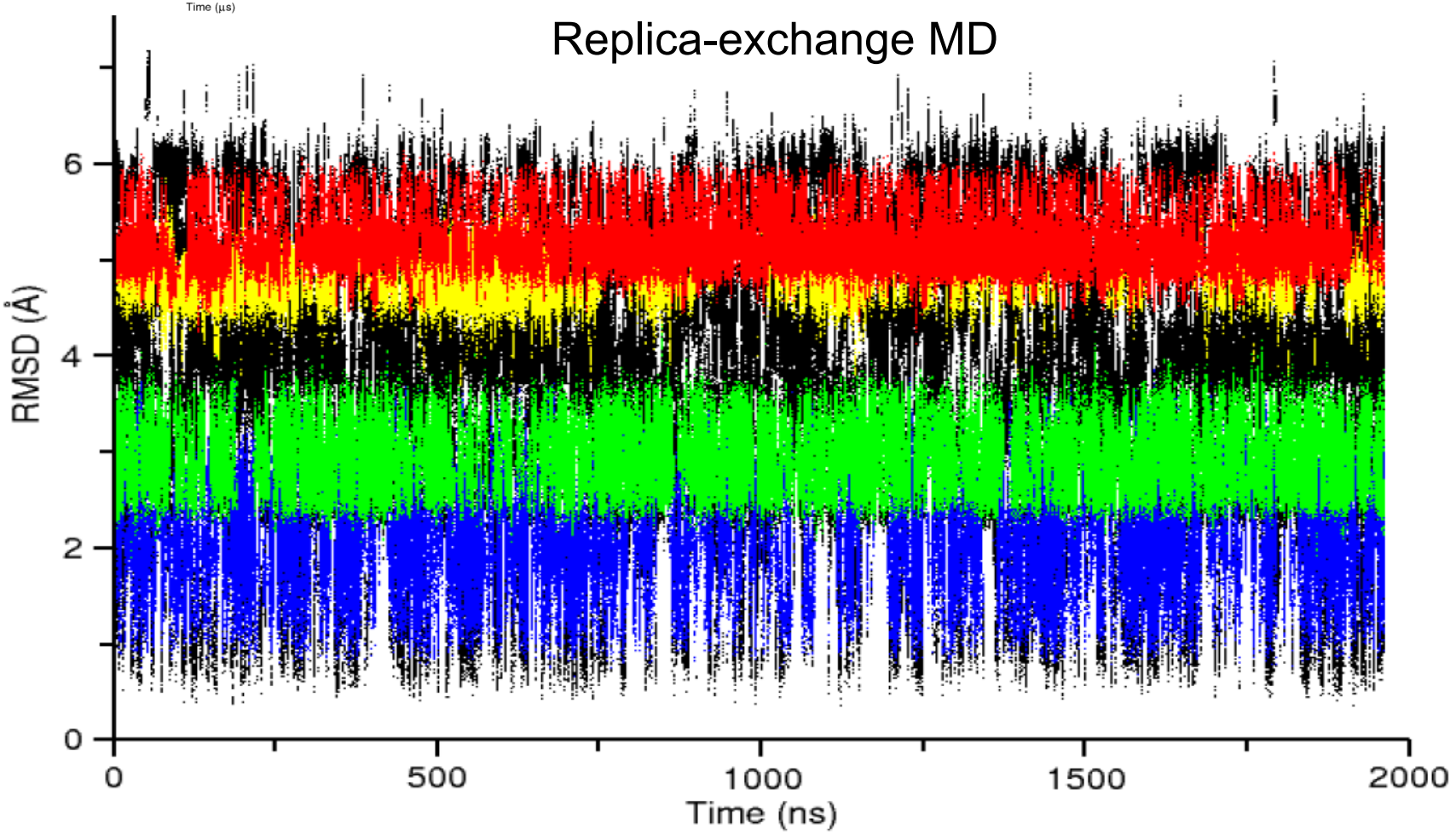
Current players: Cheatham, Roitberg, Simmerling, York, Case

Standard MD



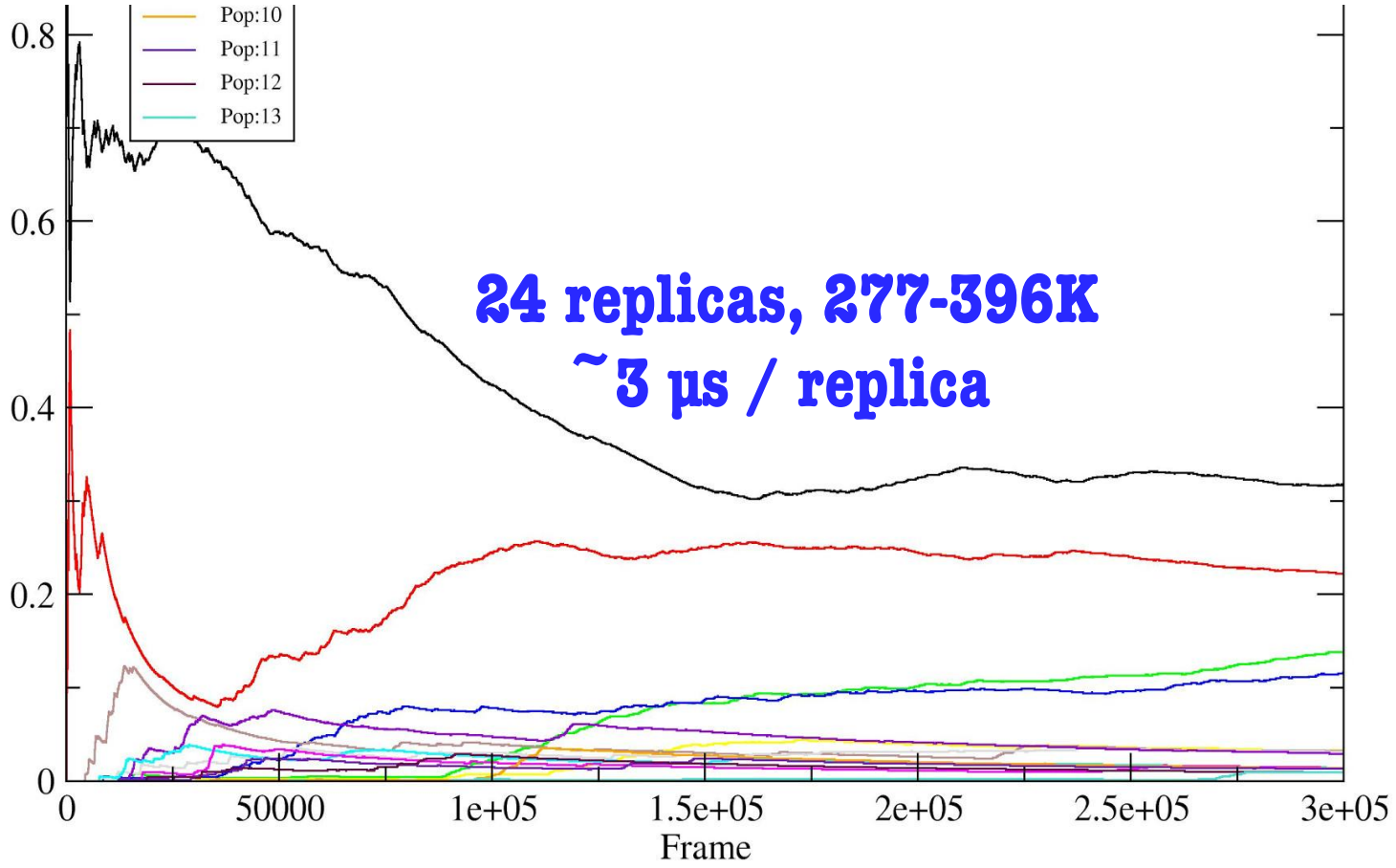
r(GACC) tetranucleotide

Replica-exchange MD



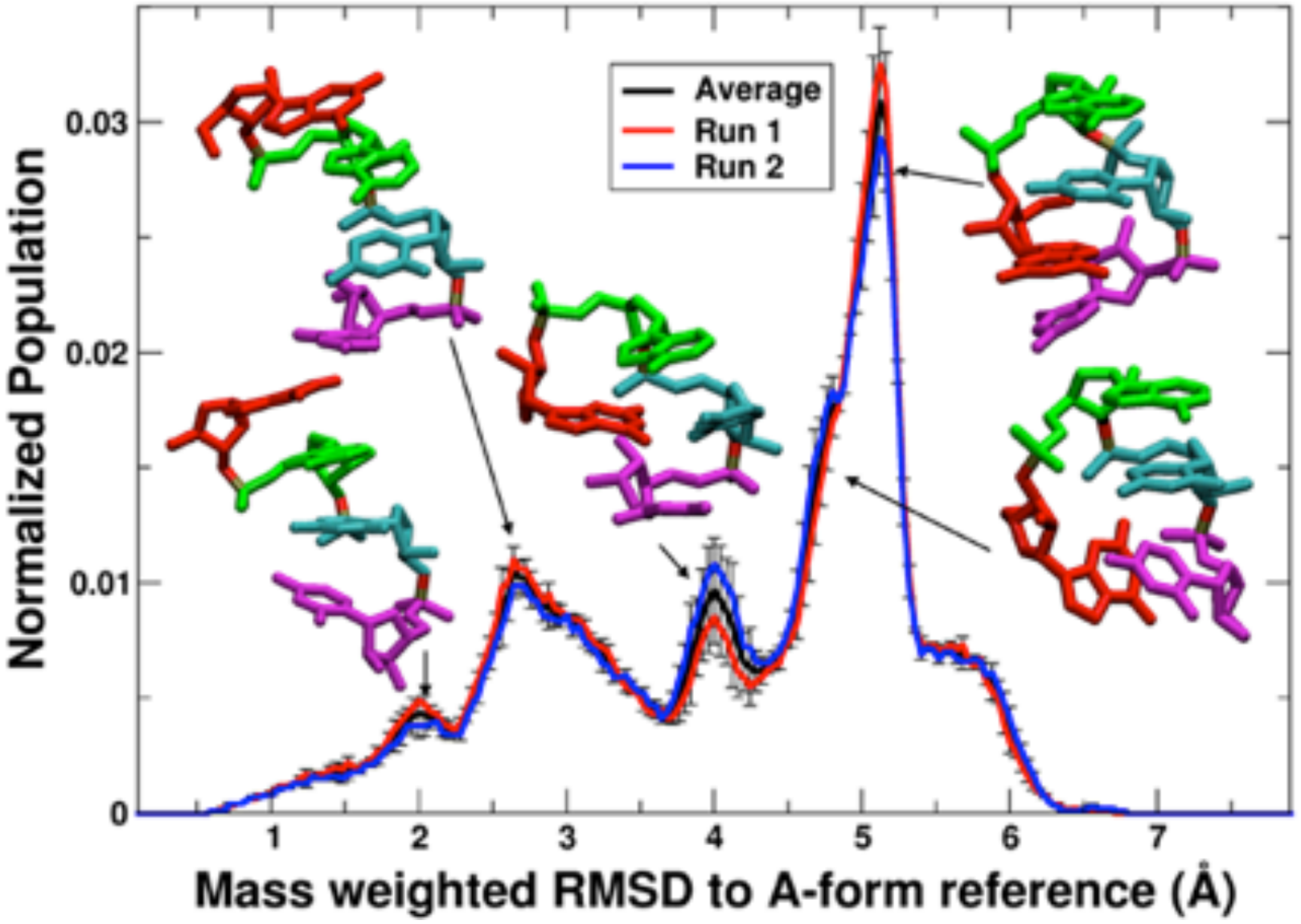
Other issues:

- **T-REMD still not "fully" converged (depending on def.)**

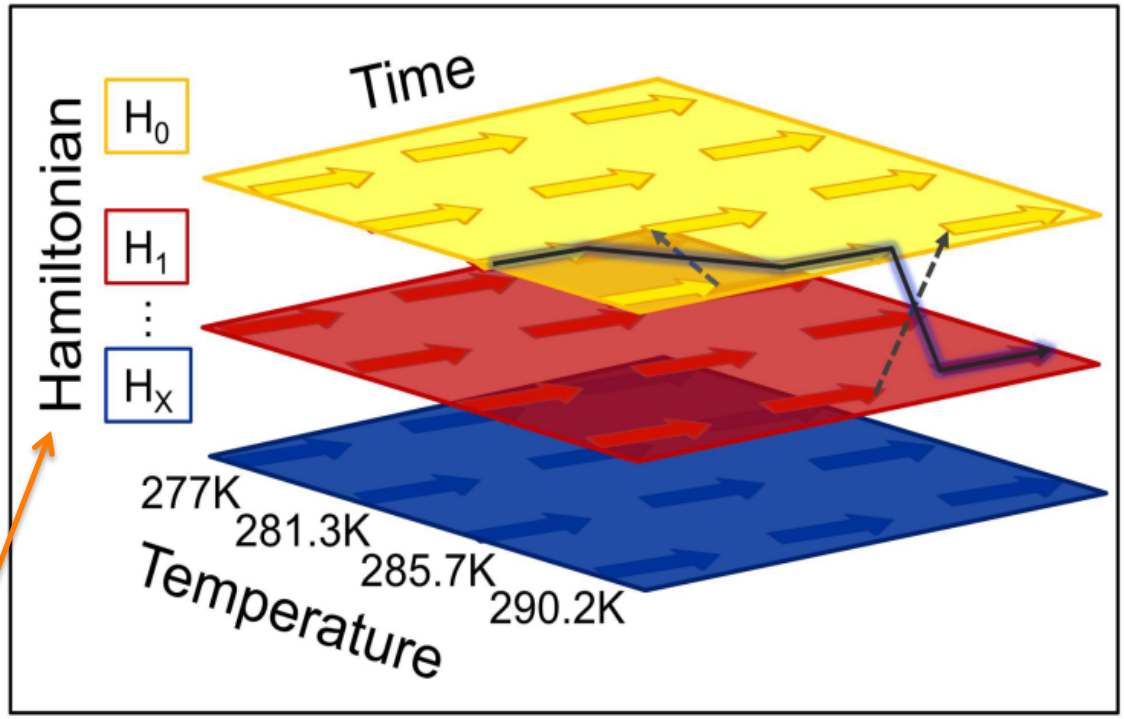


- **Not only are those four conformations populated, more like ~20+ populated > 1%**

RMSD distribution profiles: Distance from A-form reference



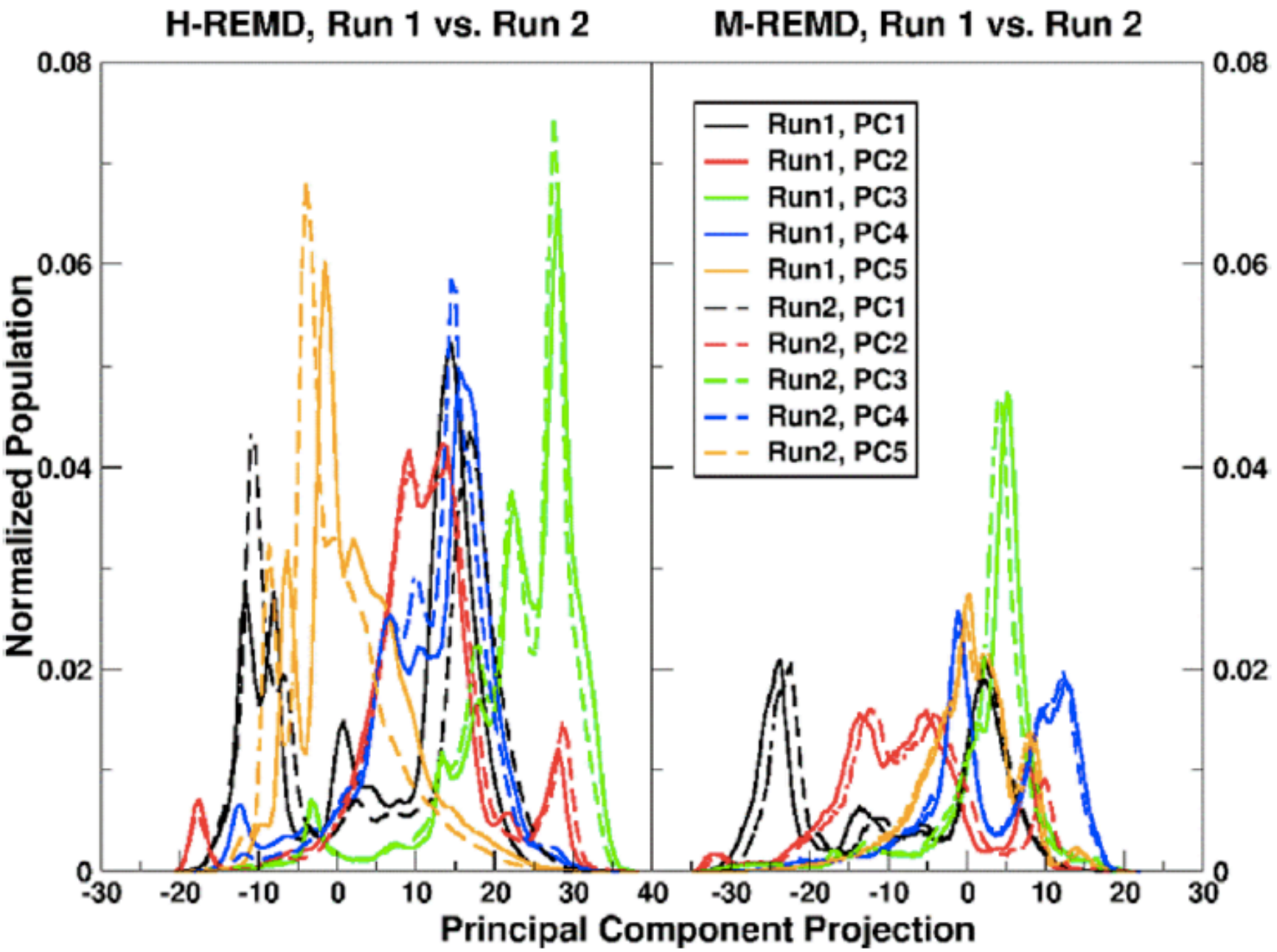
multi-D REMD – Bergonzo / Roe, Roitberg / Swails

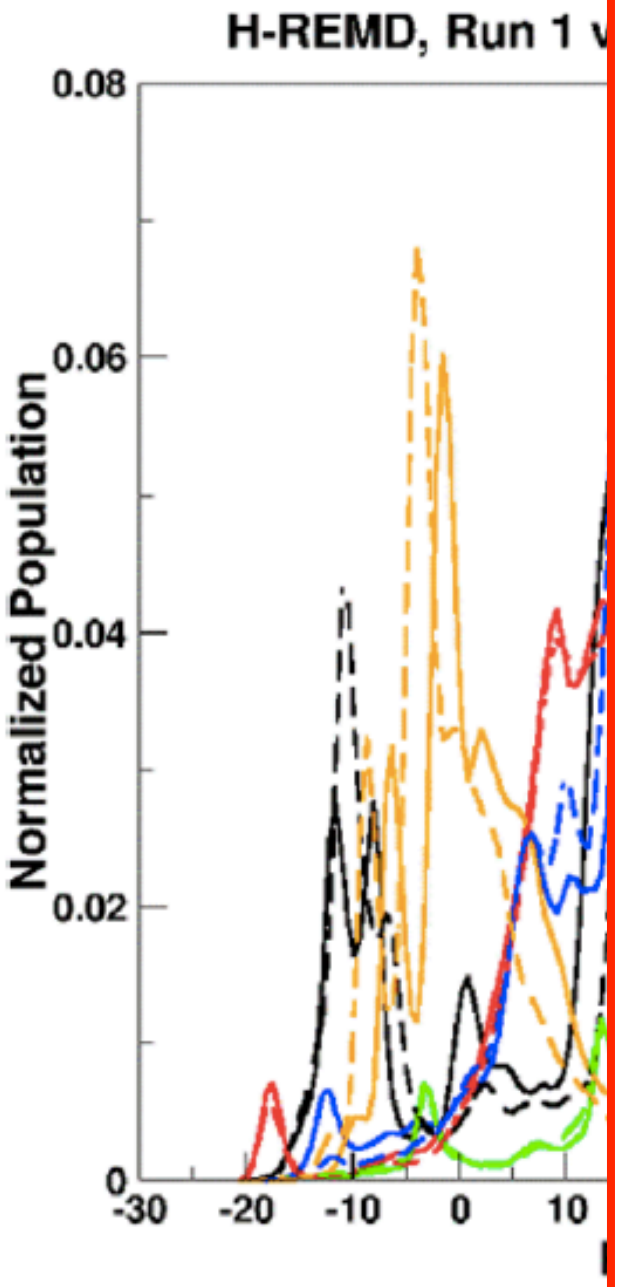


Change in "energy representation"

- **pH**
- **restraints, umbrella potentials, …**
- **force field / parameter sets**
- **biasing potentials (aMD)**

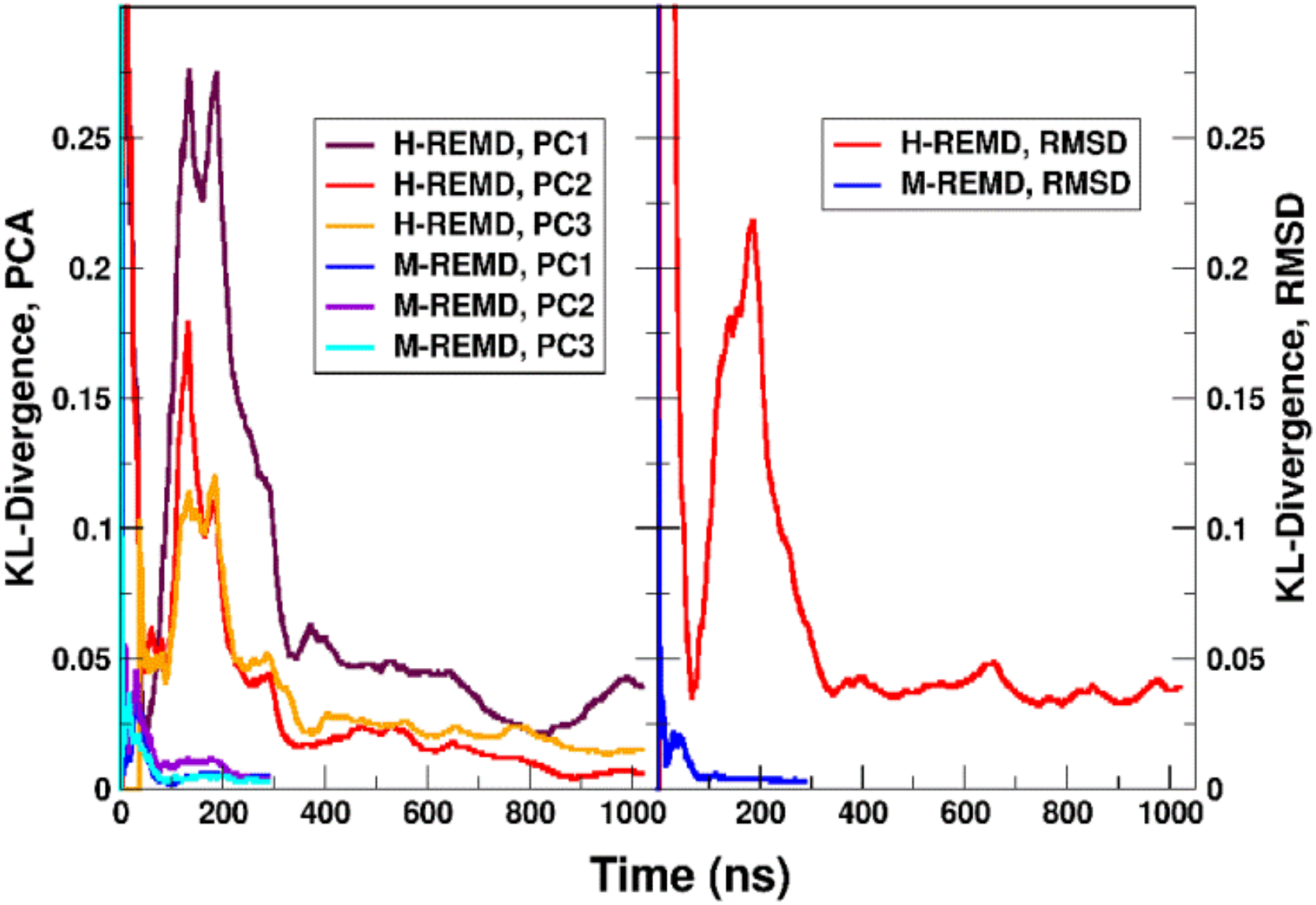
Fukunishi, H., Wanatabe, O., and Takada, S., J. Chem. Phys. 2002.
Sugita, Y., Kitao, A., and Y. Okamoto, J. Chem. Phys. 2000.

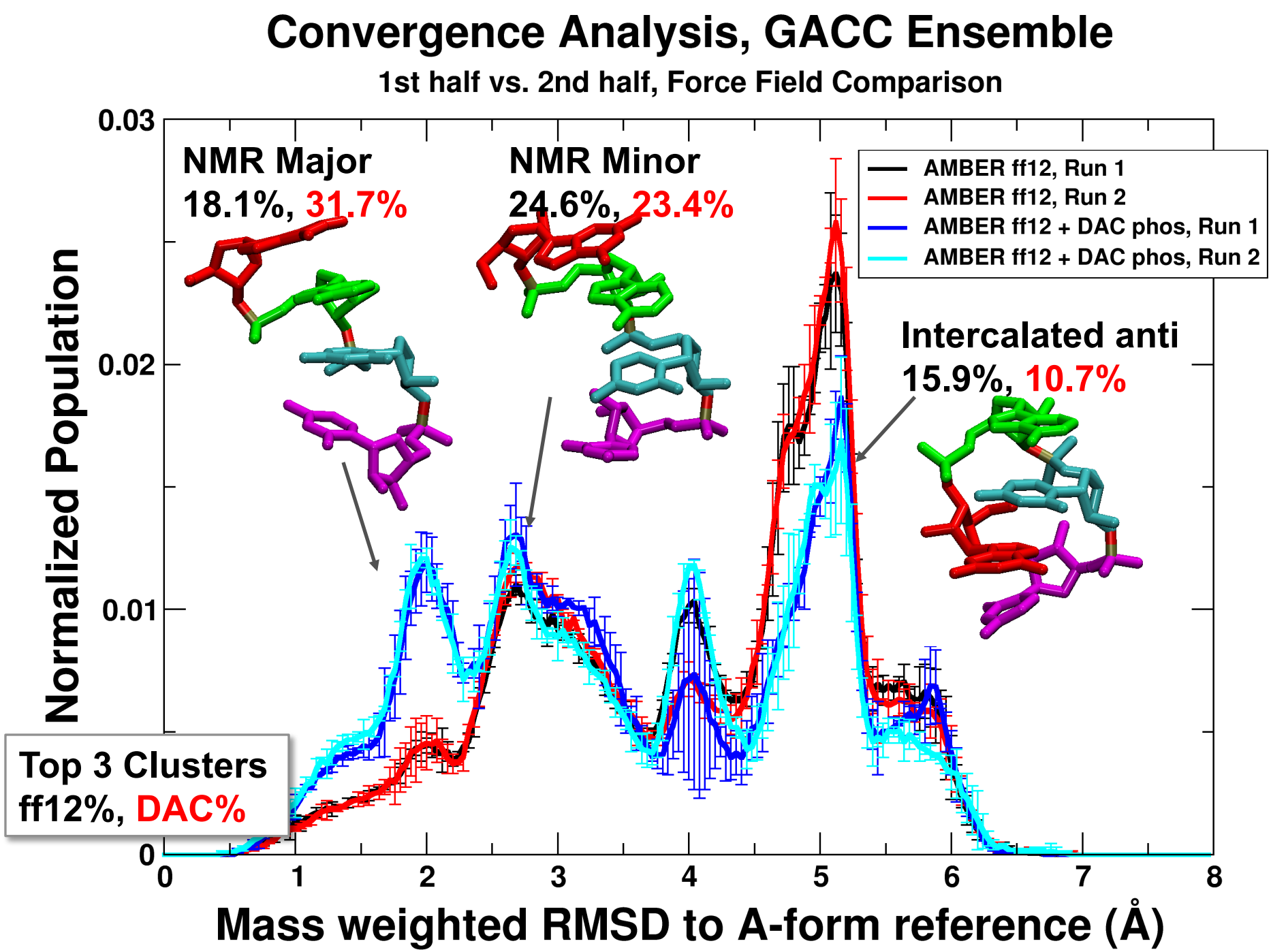

H-REMD, Run 1 vs. Run 2
M-REMD, Run 1 vs. Run 2
Normalized Population
Principal Component Projection
Run1, PC1
Run1, PC2
Run1, PC3
Run1, PC4
Run1, PC5
Run2, PC1
Run2, PC2
Run2, PC3
Run2, PC4
Run2, PC5
0.08
0.06
0.04
0.02
0
-30
-20
-10
0
10
20
30
40

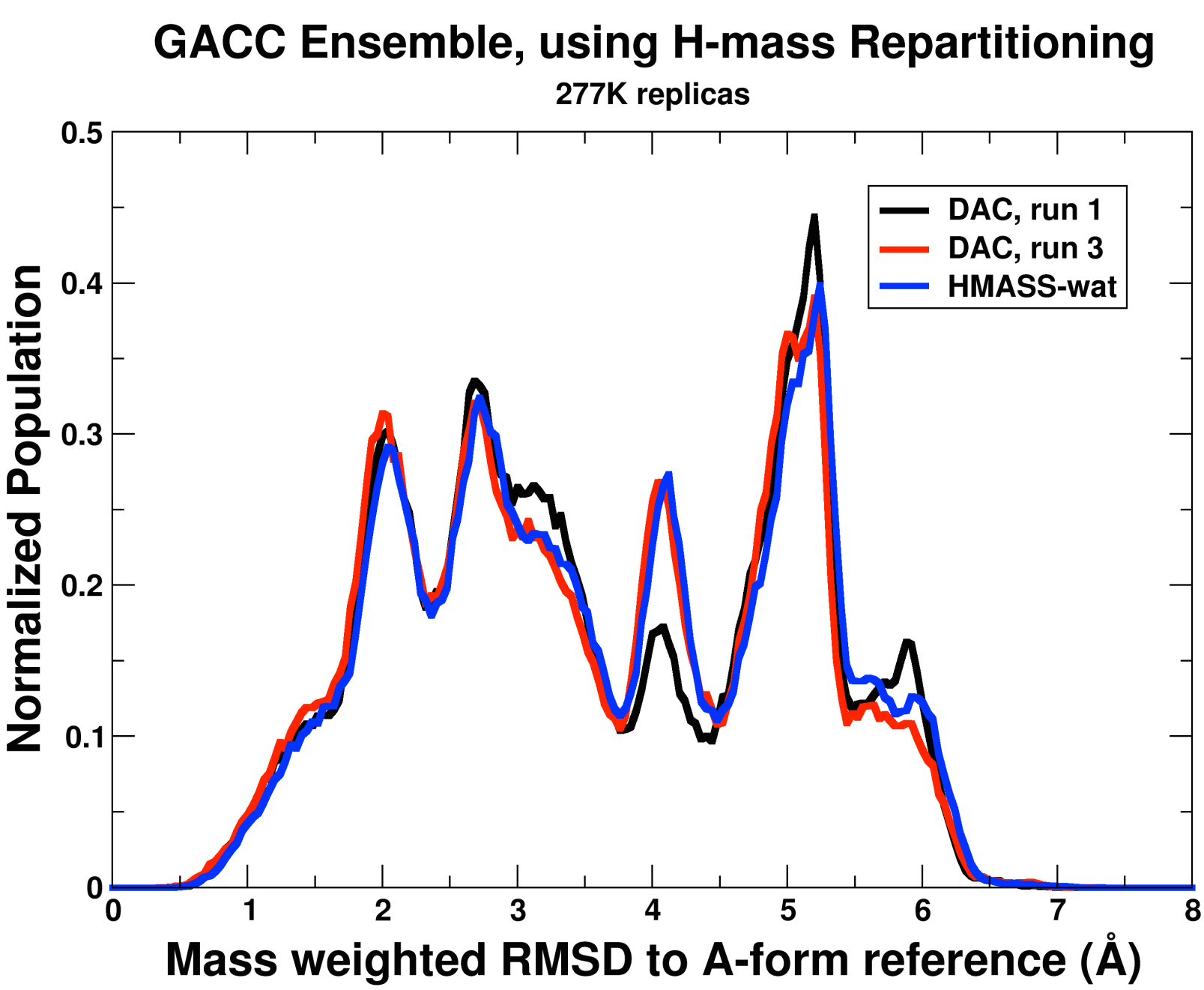


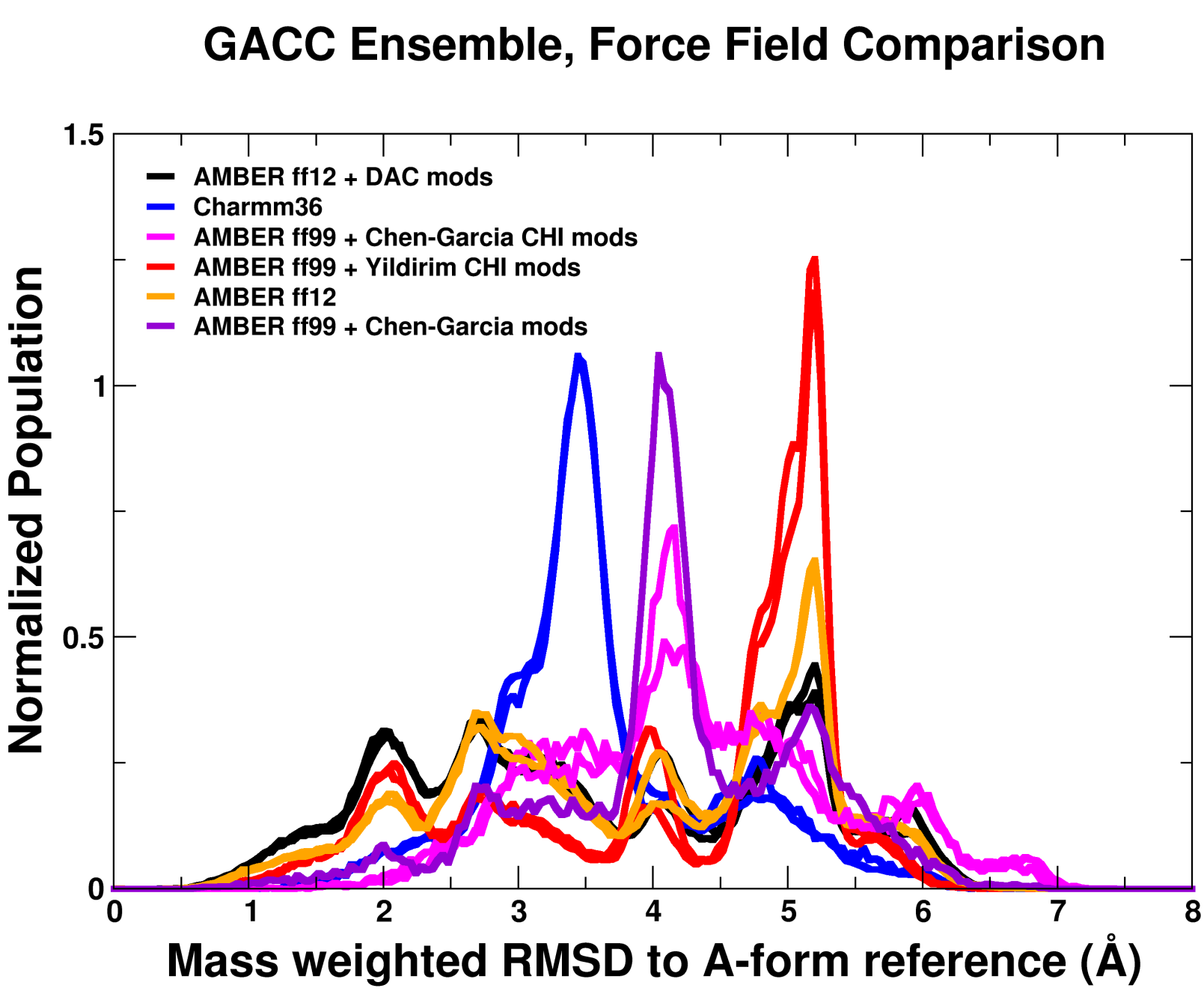
CPPTRAJ in AmberTools

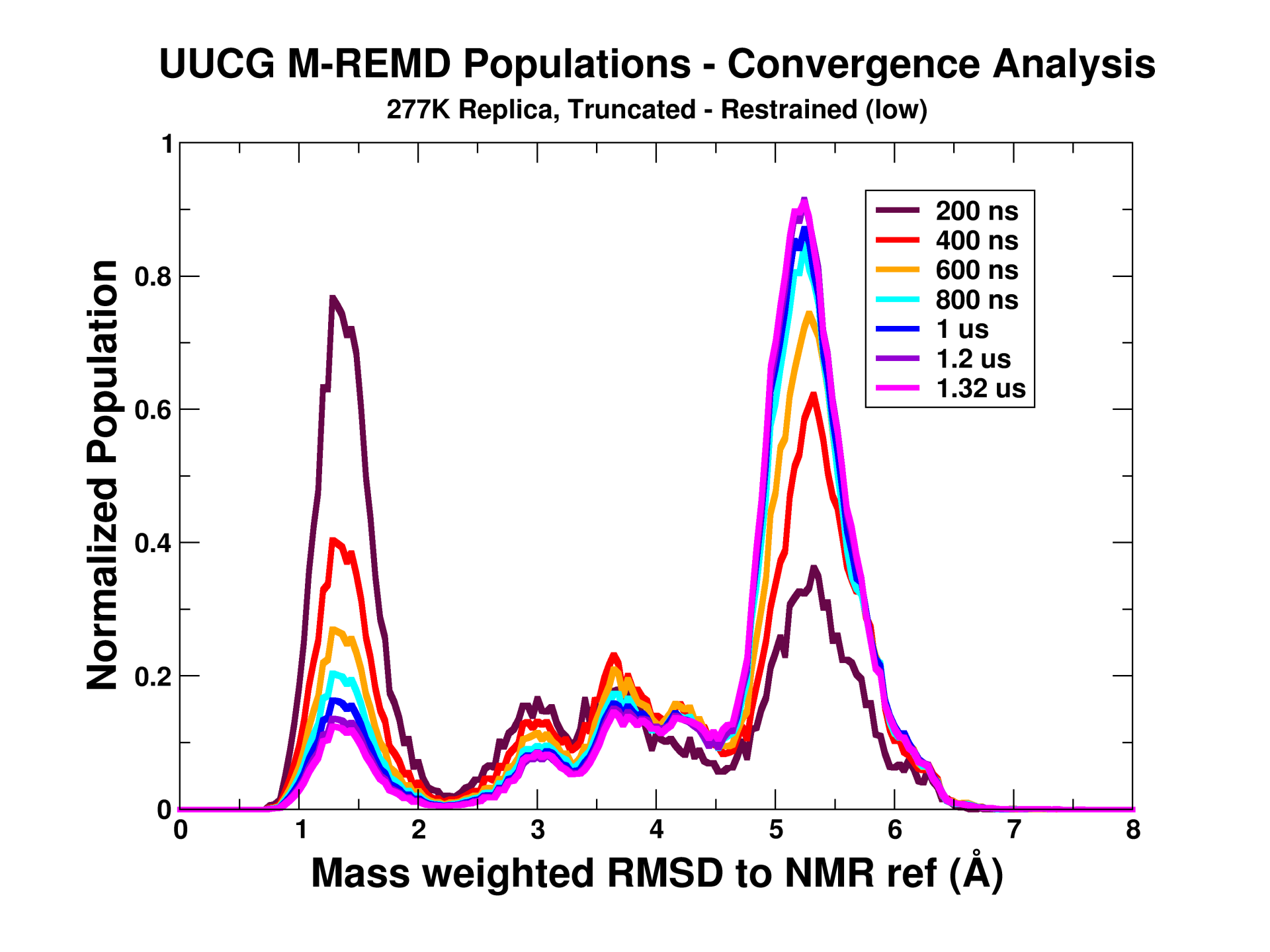
```
# Read in both trajectories
#
trajin traj.run1.nc
trajin traj.run2.nc
# RMS-fit to first frame
#
rms first :1-4&!@H=
# Create an average structure
#
average gaccAvg.rst7 ncrestart
# Save coordinates as 'crd1'
#
createcrd crd1
run
# Fit to average structure
#
reference gaccAvg.rst7.1 [avg]
# RMS-fit to average structure
#
crdaction crd1 rms ref [avg] :1-4&!@H=
# Calculate coordinate covariance matrix
#
crdaction crd1 matrix covar :1-4&!@H= name gaccCovar
# Diagonalize coordinate covariance matrix, first 15 E.vecs
#
runanalysis diagmatrix gaccCovar out evecs.dat vecs 15
# Now create separate projections for each trajectory
#
crdaction crd1 projection P1 modes evecs.dat \
    beg 1 end 15 :1-4&!@H= crdframes 1,$STOP1
crdaction crd1 projection P2 modes evecs.dat \
    beg 1 end 15 :1-4&!@H= crdframes $START2,last
# Now histogram first 5 projections for each
#
hist P1:1,*,*,*,100 out pca.hist.agr norm name P1-1
hist P1:2,*,*,*,100 out pca.hist.agr norm name P1-2
```

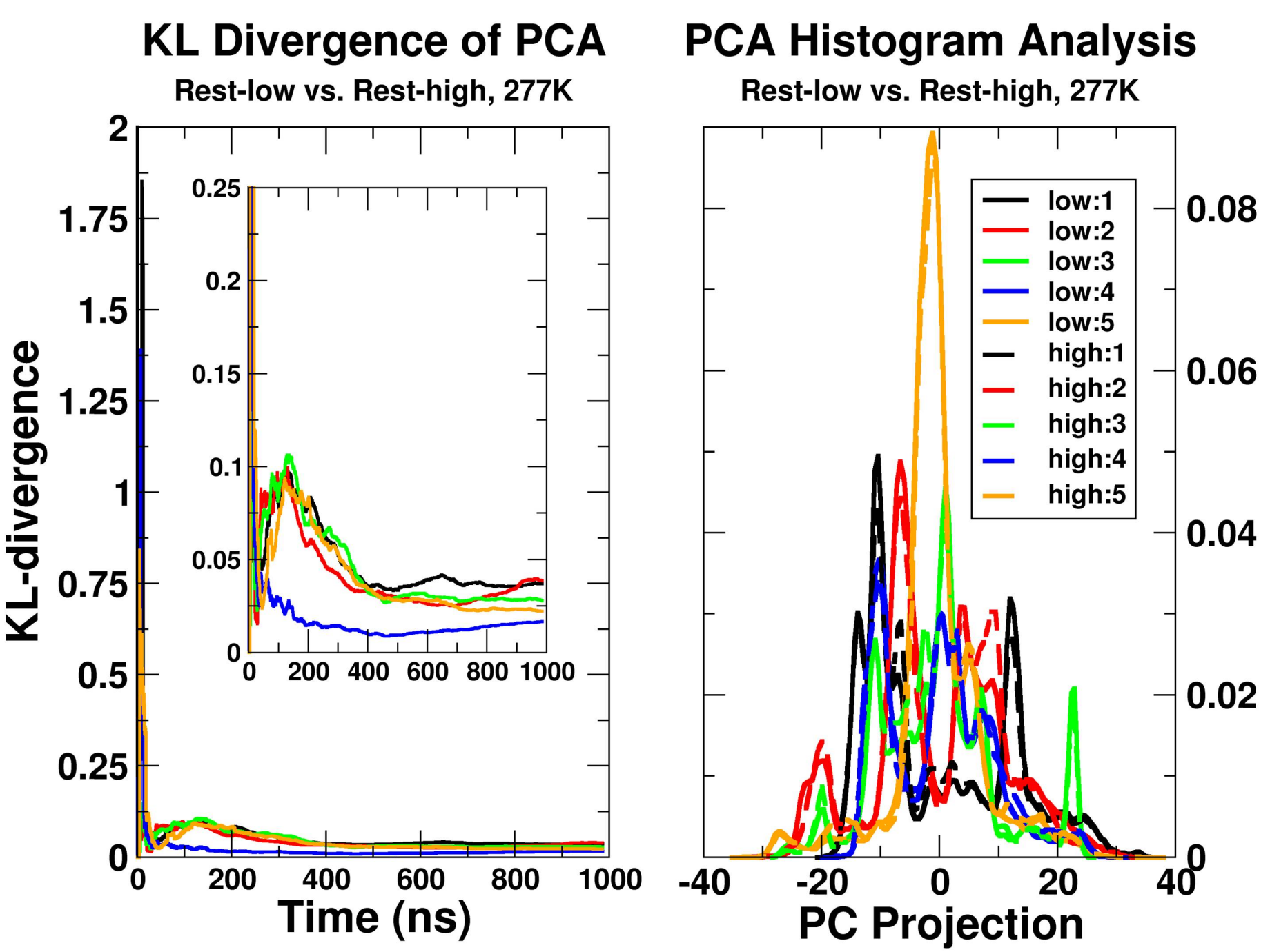

H-REMD, PC1
H-REMD, PC2
H-REMD, PC3
M-REMD, PC1
M-REMD, PC2
M-REMD, PC3
H-REMD, RMSD
M-REMD, RMSD
KL-Divergence, PCA
KL-Divergence, RMSD
Time (ns)
0
0.05
0.1
0.15
0.2
0.25
0
200
400
600
800
1000


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)


GACC Ensemble, using H-mass Repartitioning
277K replicas
DAC, run 1
DAC, run 3
HMASS-wat
Normalized Population
Mass weighted RMSD to A-form reference (Å)
0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7
8


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)


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


KL Divergence of PCA
Rest-low vs. Rest-high, 277K
KL-divergence
Time (ns)
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
0
200
400
600
800
1000
0.05
0.1
0.15
0.2
0.25
PCA Histogram Analysis
Rest-low vs. Rest-high, 277K
PC Projection
-40
-20
0
20
40
0.02
0.04
0.06
0.08
low:1
low:2
low:3
low:4
low:5
high:1
high:2
high:3
high:4
high:5

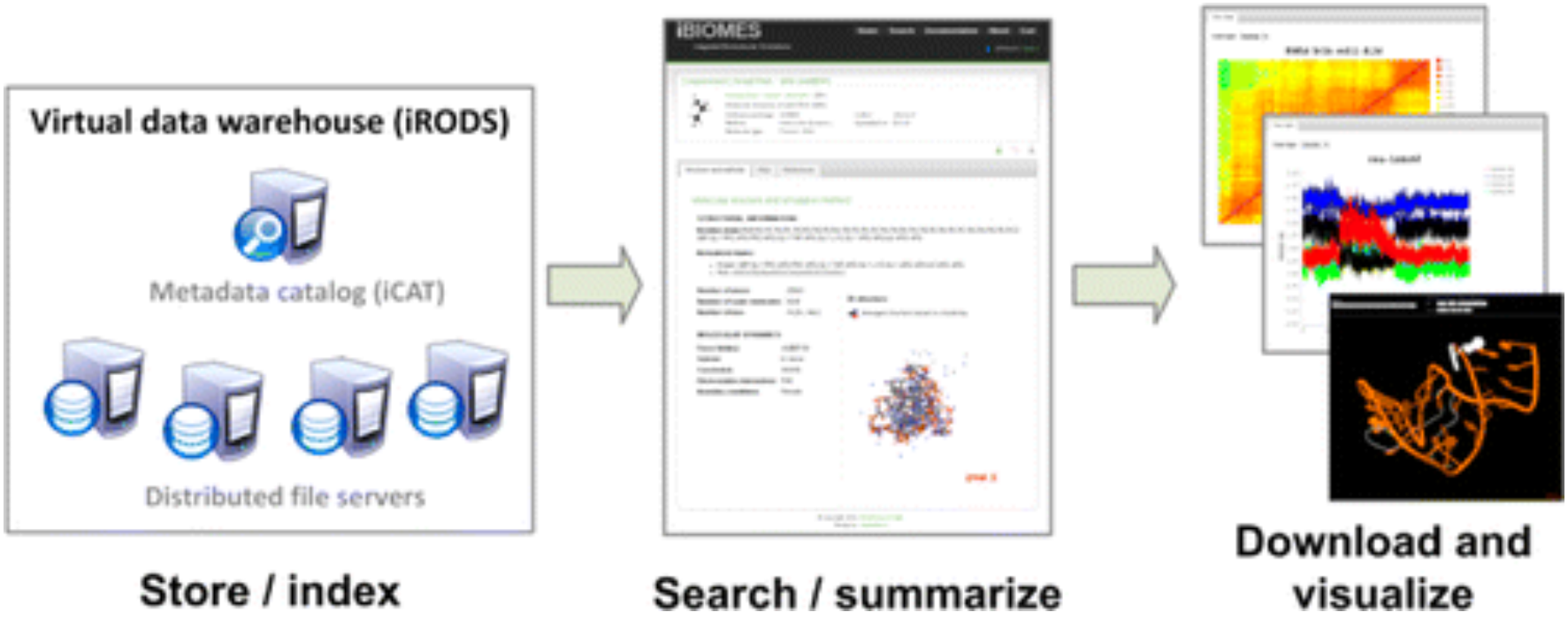
JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

iBIOMES: Managing and Sharing Biomolecular Simulation Data in a Distributed Environment

Julien C. Thibault,[†] Julio C. Facelli,[†,‡] and Thomas E. Cheatham, III*[,§]



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

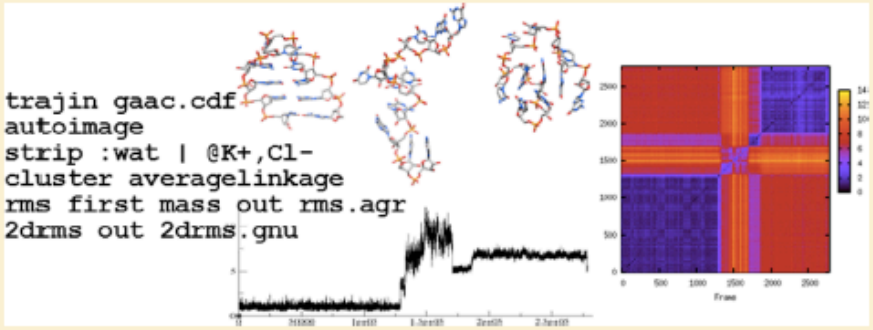
PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data

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

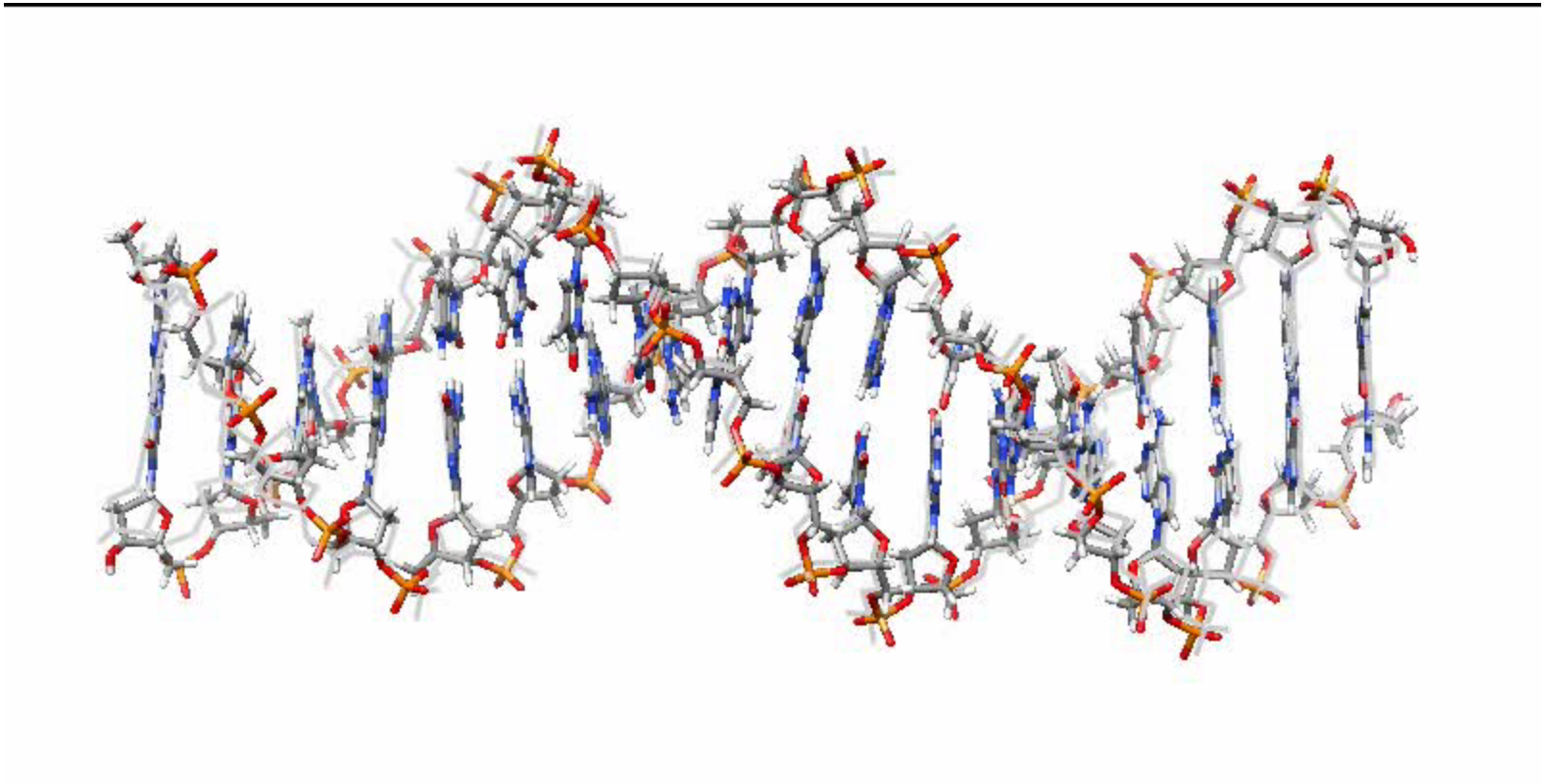
Department of Medicinal Chemistry, College of Pharmacy, 2000 South 30 East Room 105, University of Utah, Salt Lake City, Utah 84112, United States

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.



questions?



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