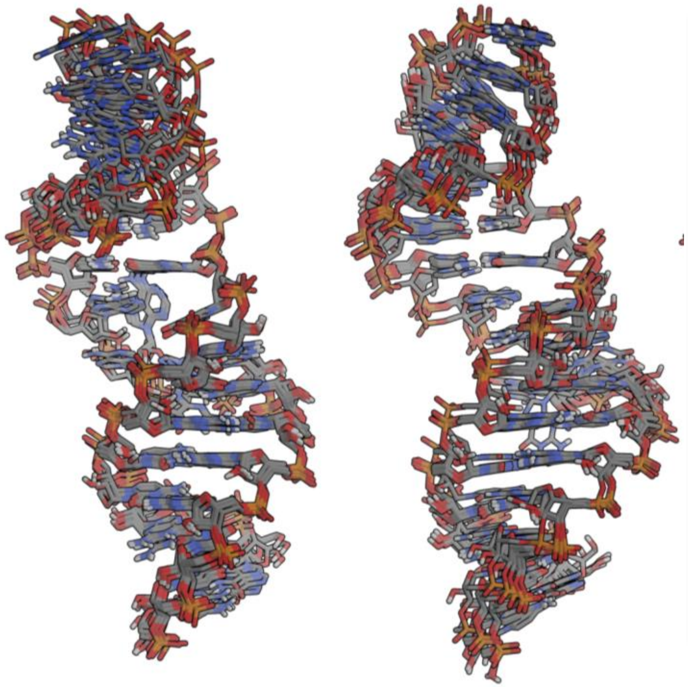
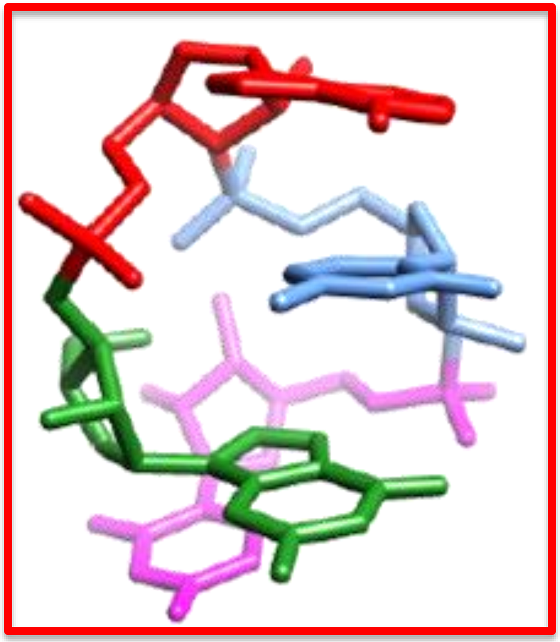
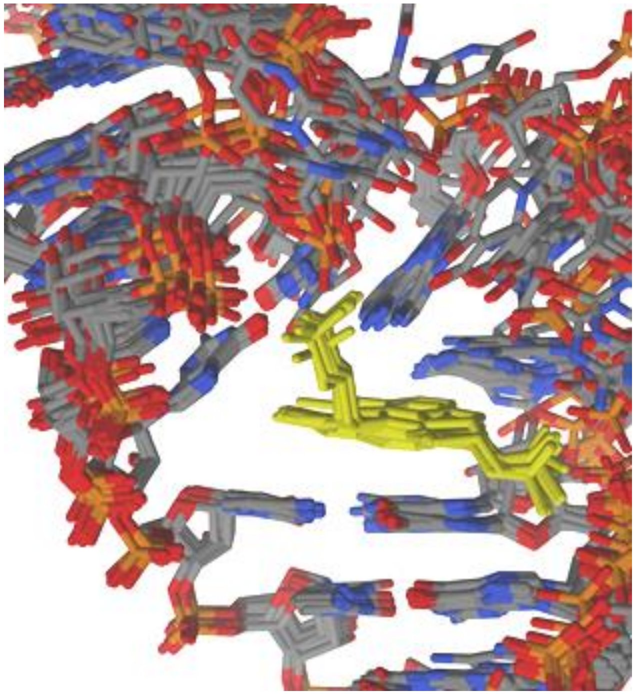
Convergence and reproducibility in molecular dynamics simulations of nucleic acids enabled by Blue Waters







Thomas E. Cheatham III
Professor, Dept. of Medicinal Chemistry, College of Pharmacy
Director, Center for High Performance Computing
University of Utah

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

$$$:

National Science Foundation
WHERE DISCOVERIES BEGIN

NIH R01-GM098102	"RNA-ligand interactions: simulation & experiment"
NSF CHE-1266307	"CDS&E: Tools to facilitate deeper data analysis, …"
NSF ACI-1521728	"RAPID: Optimizing … Ebola membrane fusion inhibitor … design"
NSF ACI-1443054	"CIF21 DIBBS: Middleware and high performance analytics…"
NSF ACI-1341034	"CC-NIE Integration: Science slices…" network DMZ
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
~12M core hours

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

~7-14M GPU hours
!!!

CHPC

~3M hours

Accurate modeling of RNA and other biomolecules r

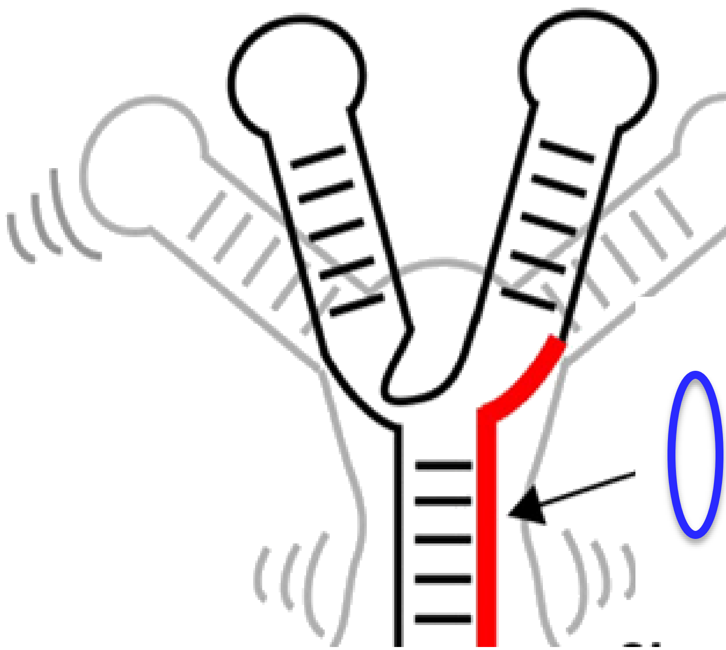
accurate and fast simulation methods

validated RNA, protein, water, ion, and ligand “forc

“good” experiments to assess results

dynamics and complete sampling: (convergence, repr

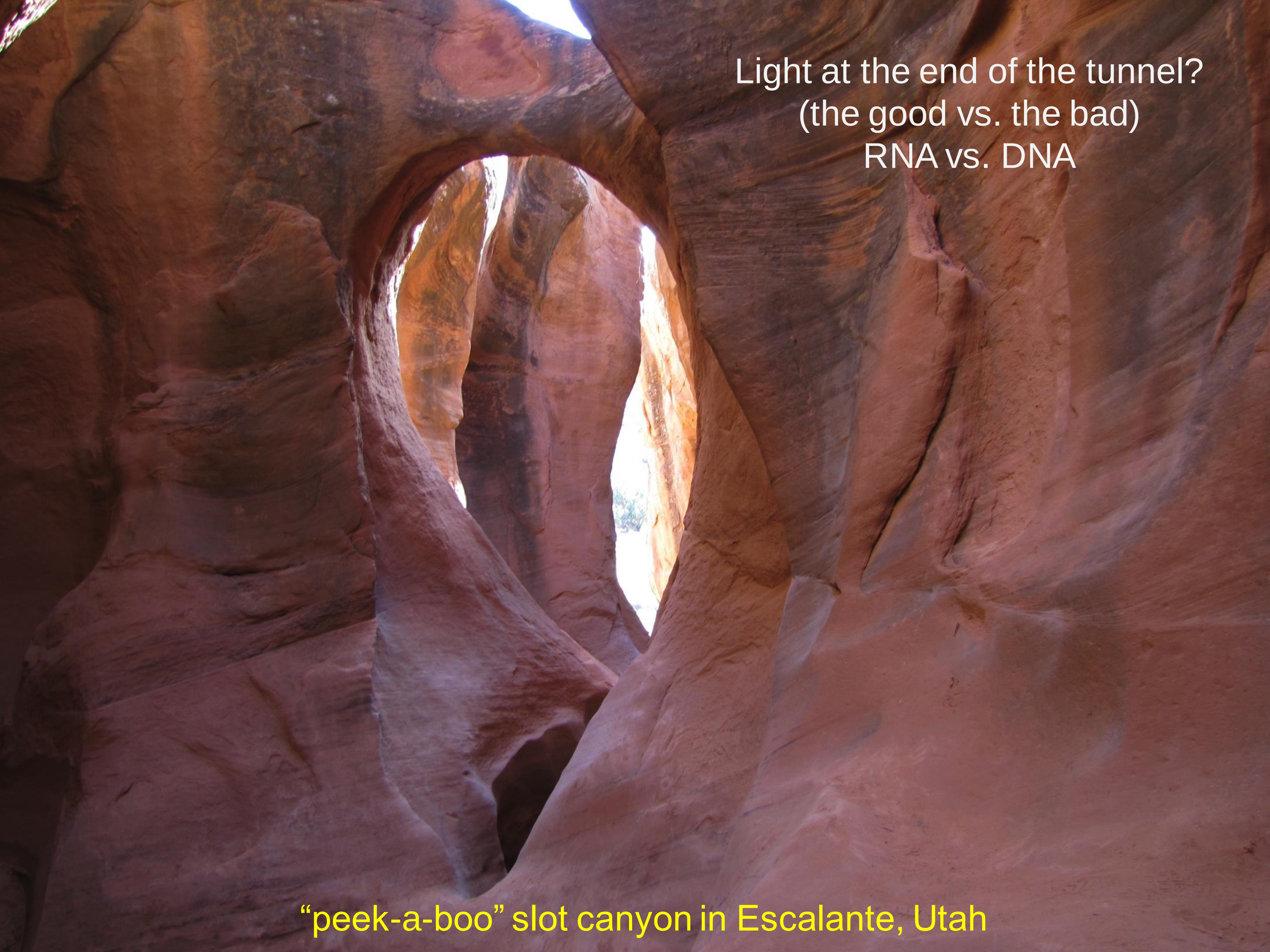
Question: Is the movement real or artifact?



conformational selectio

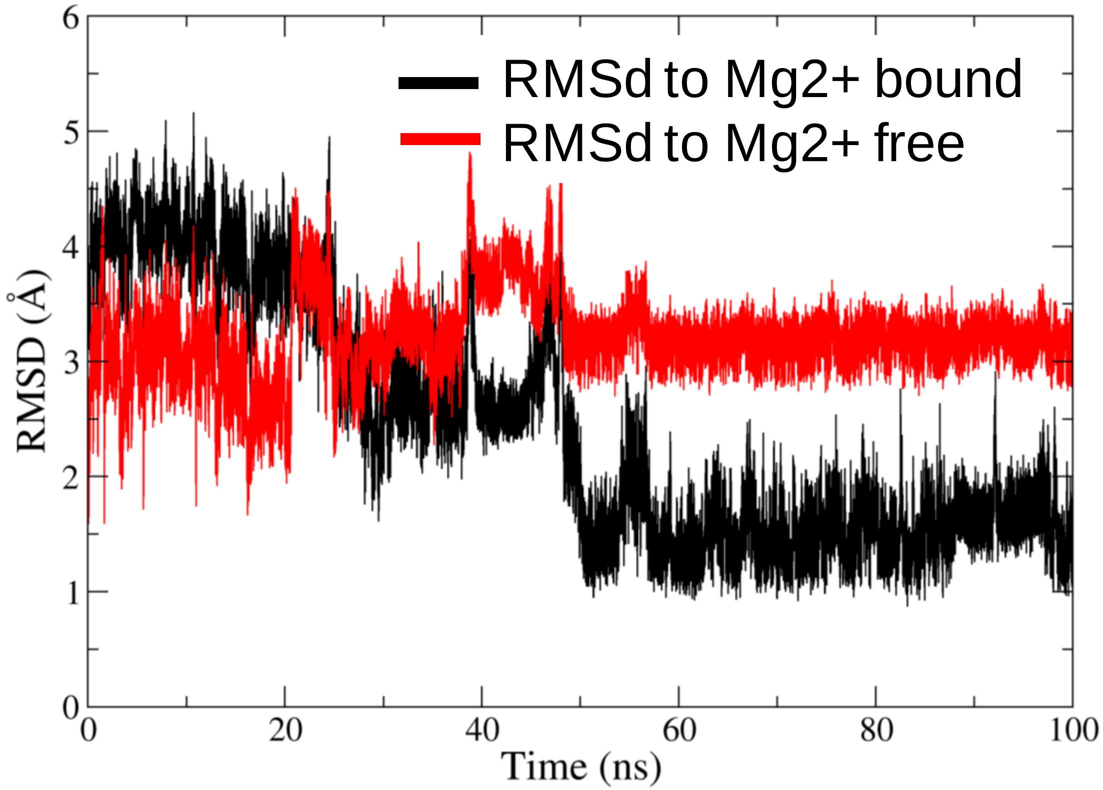
vs.

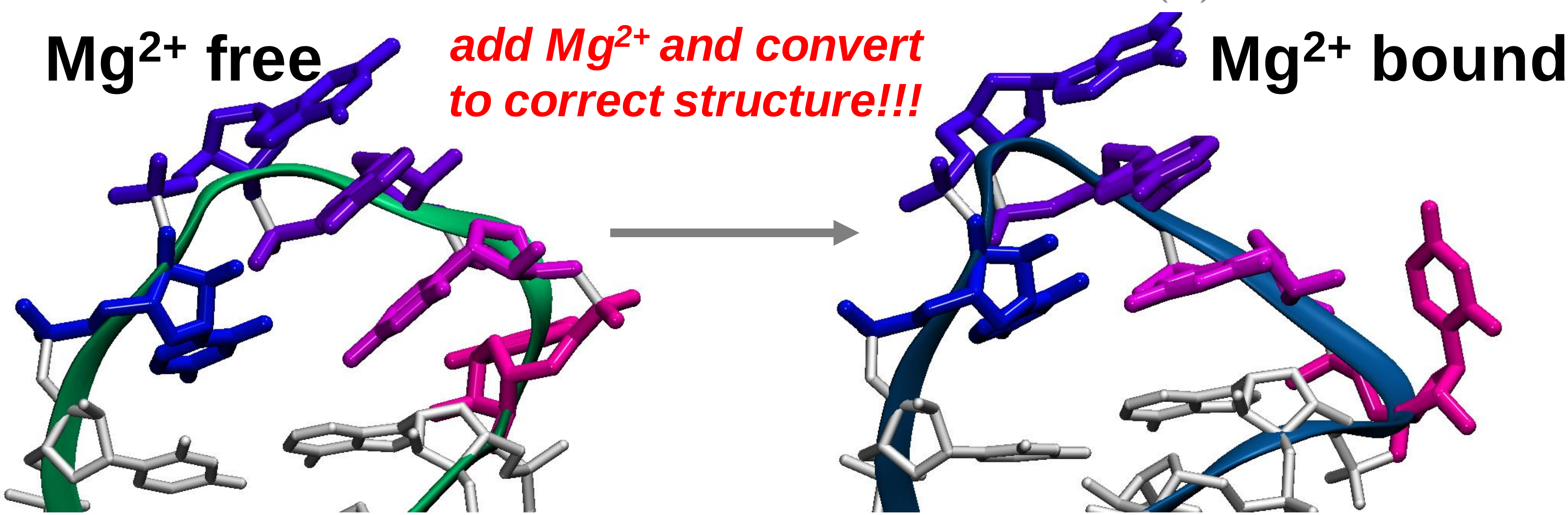
induced fit


Light at the end of the tunnel?
(the good vs. the bad)
RNA vs. DNA
“peek-a-boo” slot canyon in Escalante, Utah

We're seeing some progress!!!

(vsrSL5)





~1978 - present

amber

Assisted Model Building with Energy Refinement

code vs. force field

late 60's: CFF (consistent force field) + early code
{Warshel, Levitt, Lifson}

first protein simulation ~1975

1978: Bruce Gelin thesis @ Harvard {Karplus}

GROMOS

Amber 1.1, **1981**
(minimization only, f'')

CHARMM

ENCAD

Discover

GROMACS

Amber 2, **1984**
(+ dynamics)

NAMD

first nucleic acid simulation in H_2O ~1985

~1978 - present

amber

Assisted Model Building with Energy Refinement

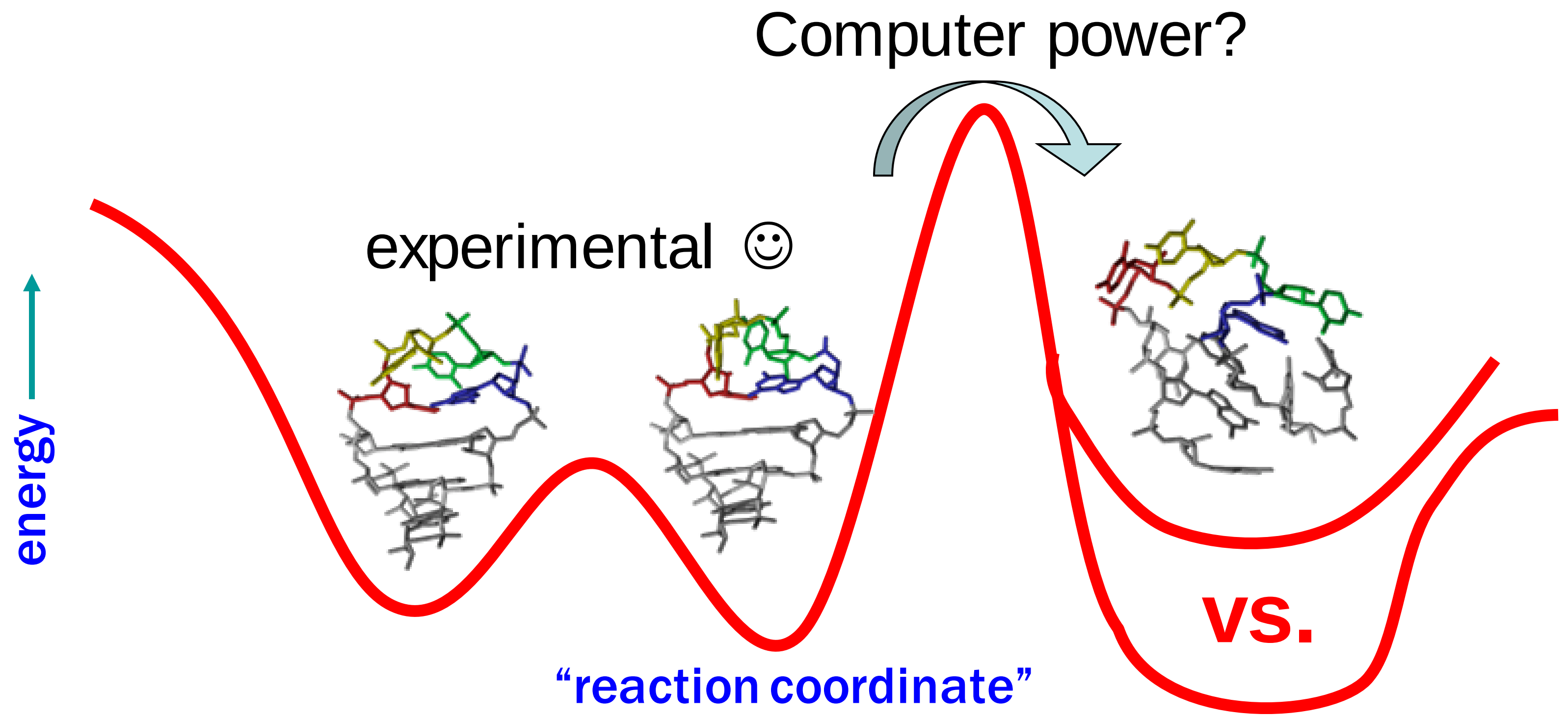
code vs. force field

Amber 14 released April, 2014; AmberTools 15, May 2015

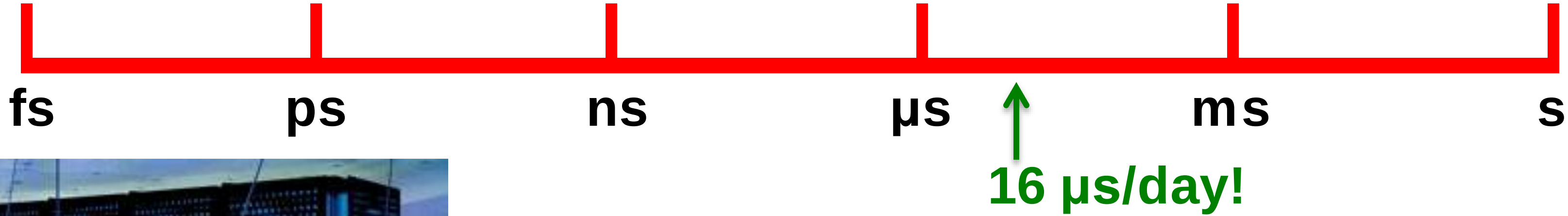
- 1.23x increase in GPU performance
  [fully deterministic, mixed SP/fixed precision, ||-ized]
- support for M-REMD simulation and analysis
- constant pH
- new TI methods
- more methods ported to GPU
- protein ff14SB, RNA ff12, DNA ff12+χ_{OL4}+ε/ζ

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?



Simulating protein movements using Anton could aid drug design.

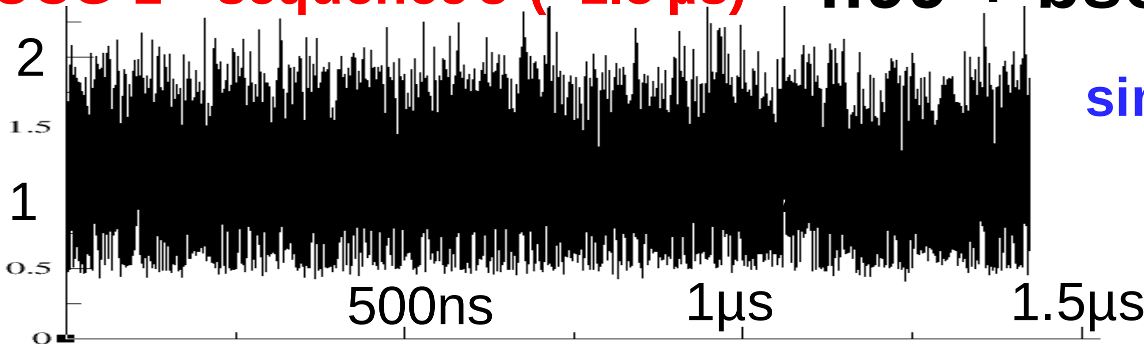
SCIENCE/AAAS

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

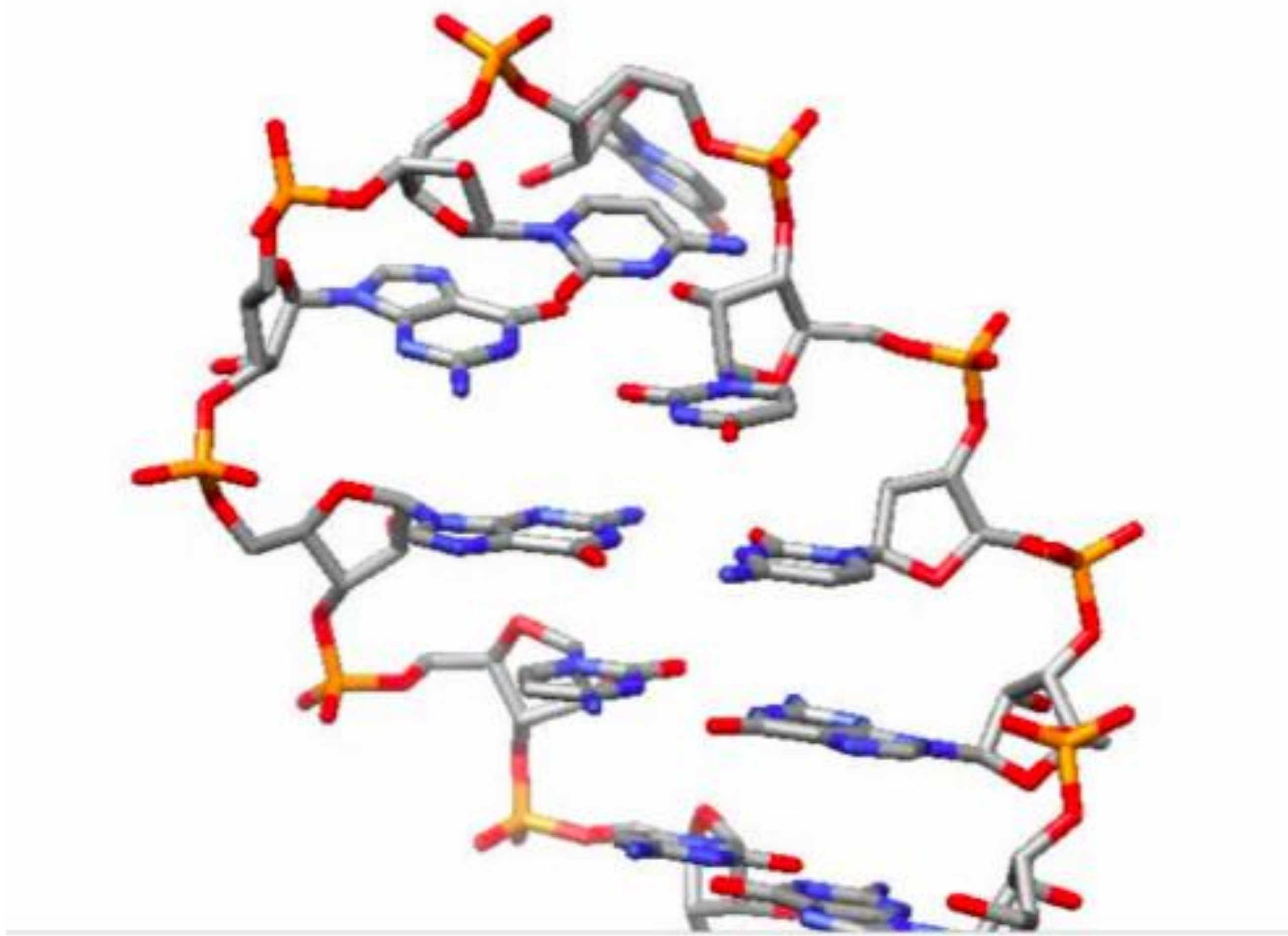
D.E. Shaw's Anton machine

UUCG-1 – sequence 3 (~1.5 μs)

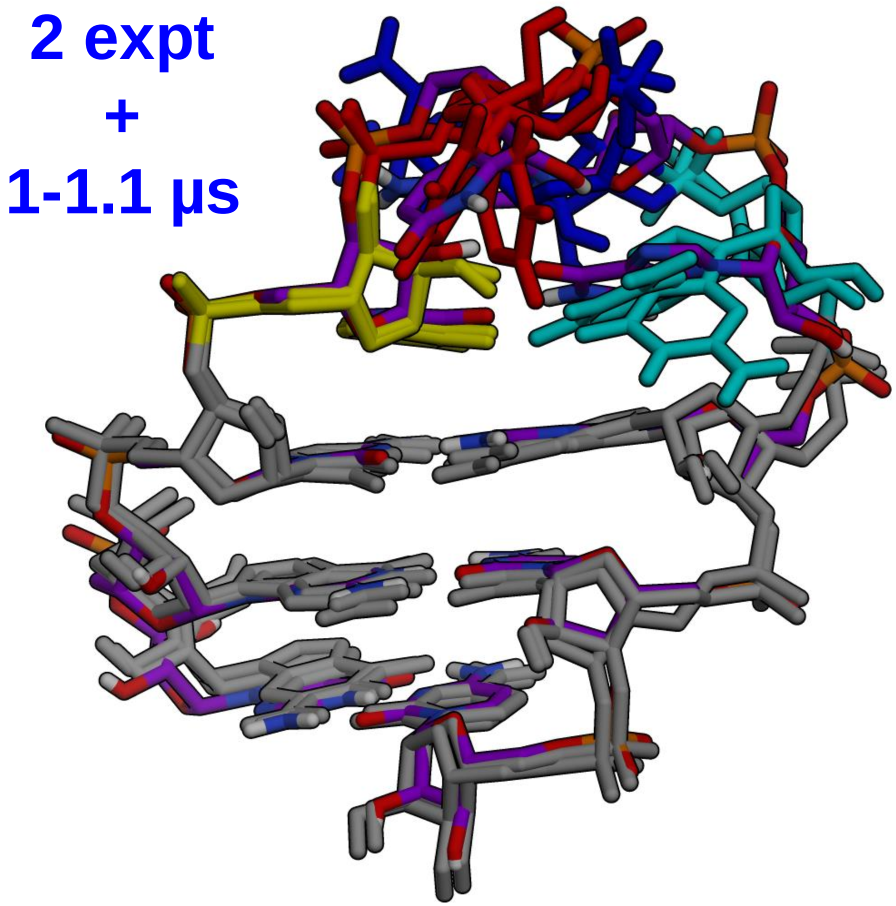
ff99 + bsc0 + OL χ fix

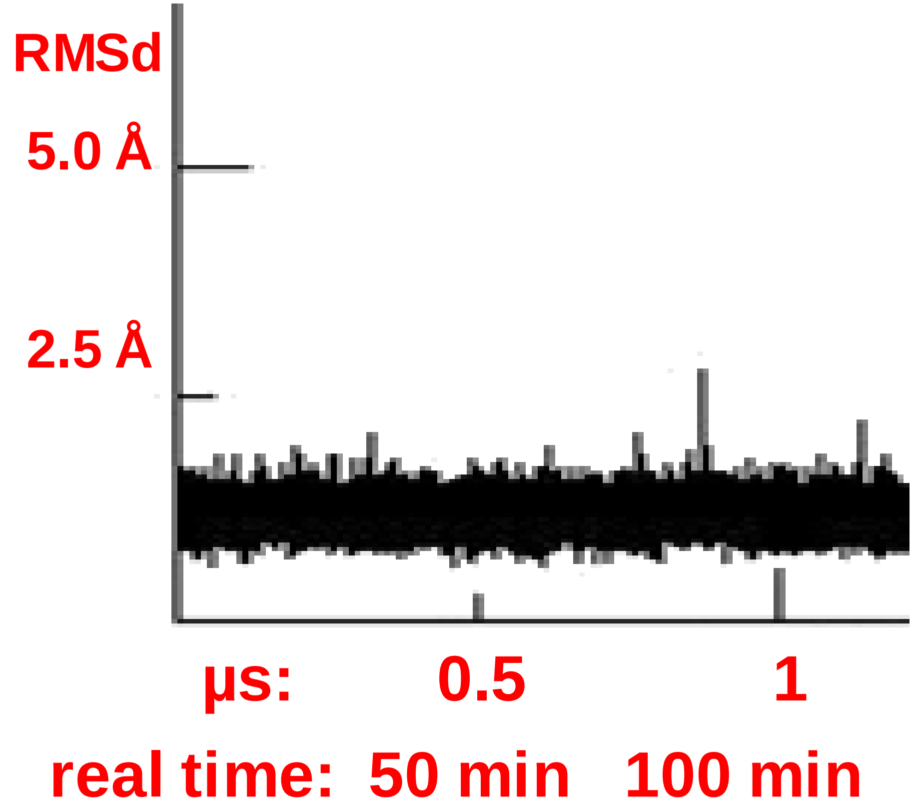


simulated w/out restraints,
modern force field,
explicit solvent



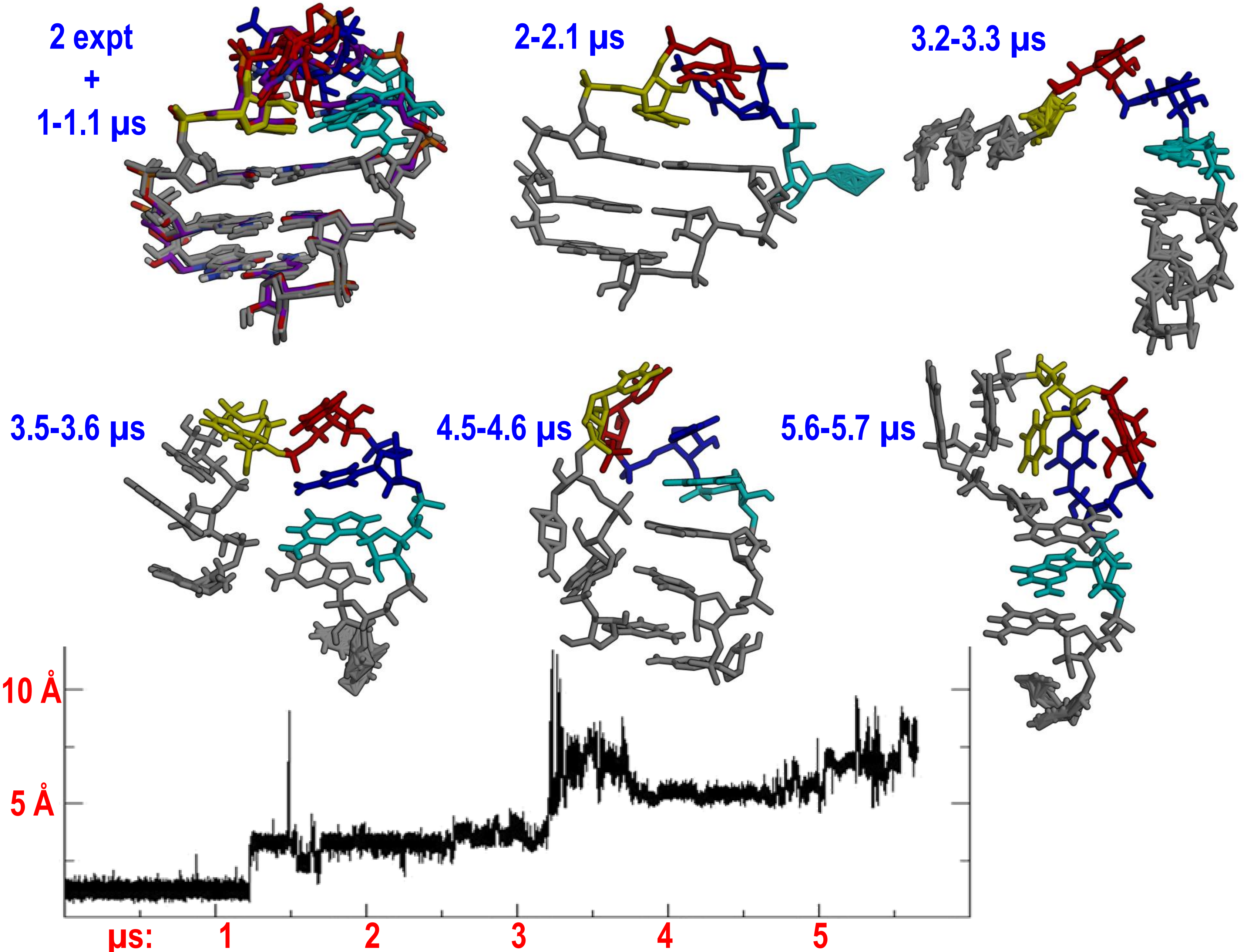
RNA UUCG tetraloop (ff99bsc0 + OL X) on Anton @ PSC:





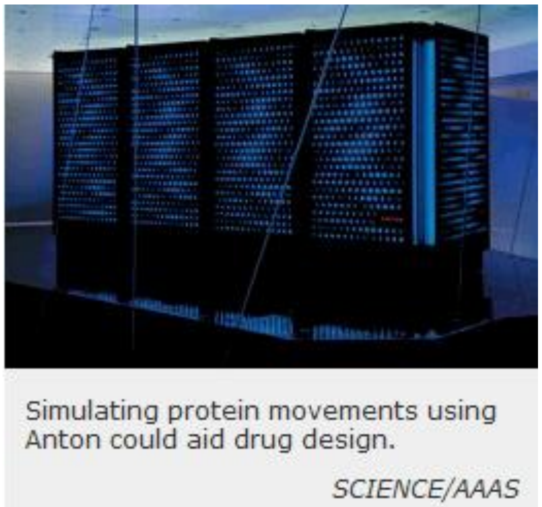
Initial tests: RNA tetraloop

RNA UUCG tetraloop (ff99bsc0 + OL X):



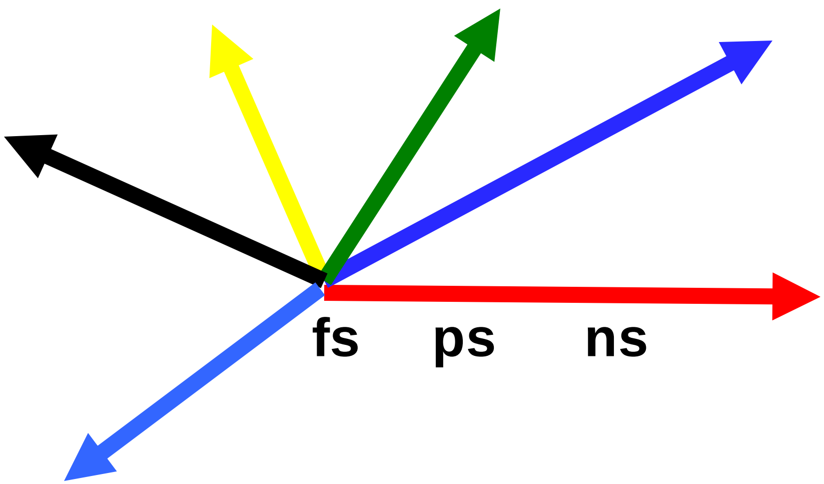
How to fully sample conformational ensemble?

16 µs/day!



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

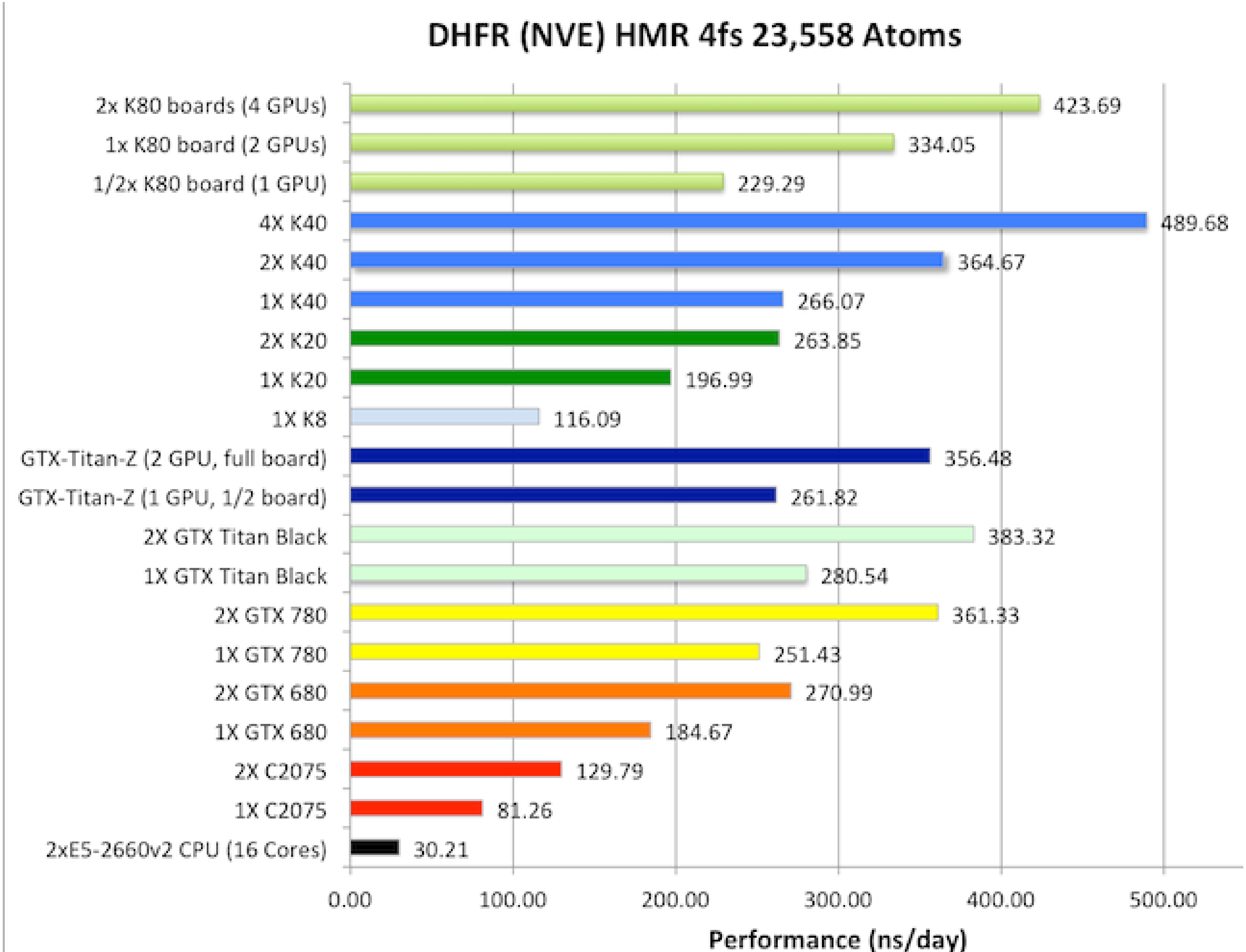
D.E. Shaw's Anton machine



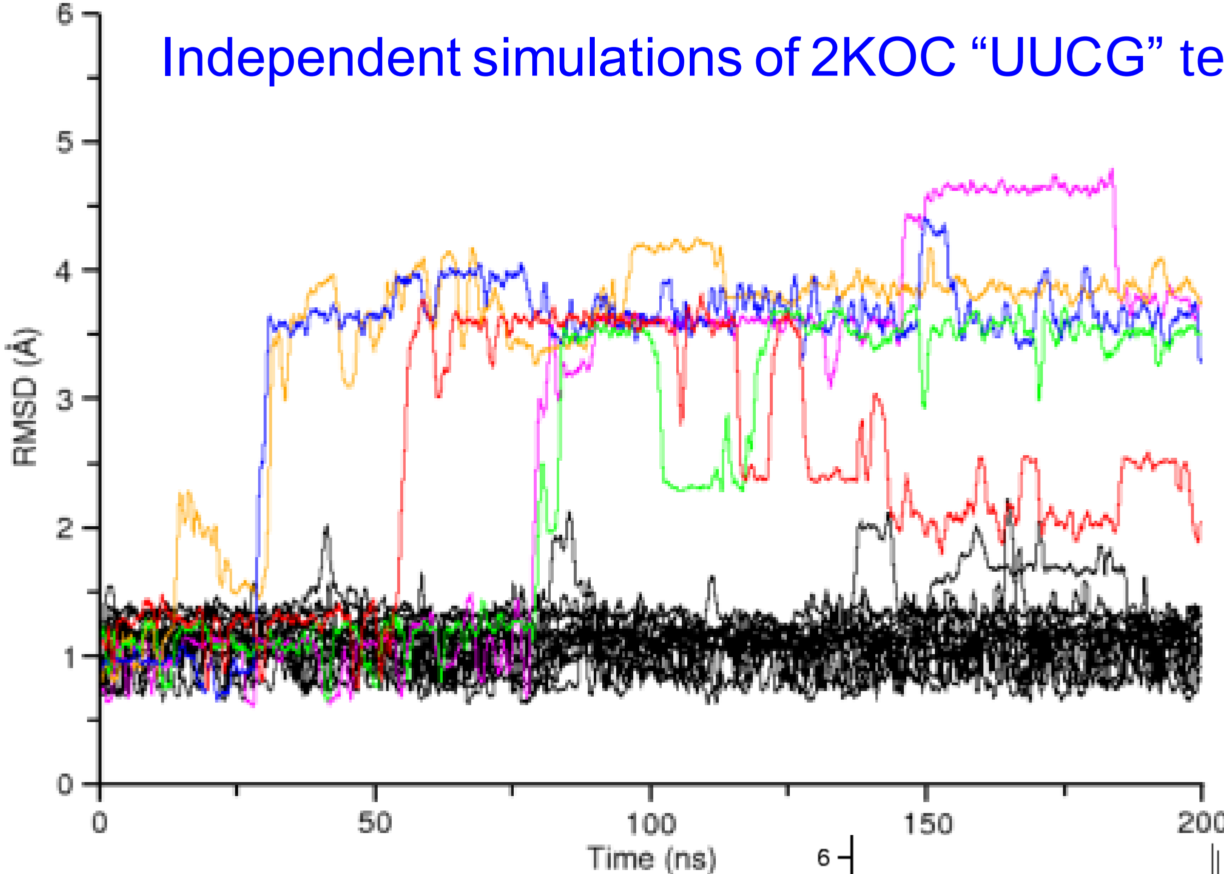
ensembles of independent simulations

AMBER on GPUs

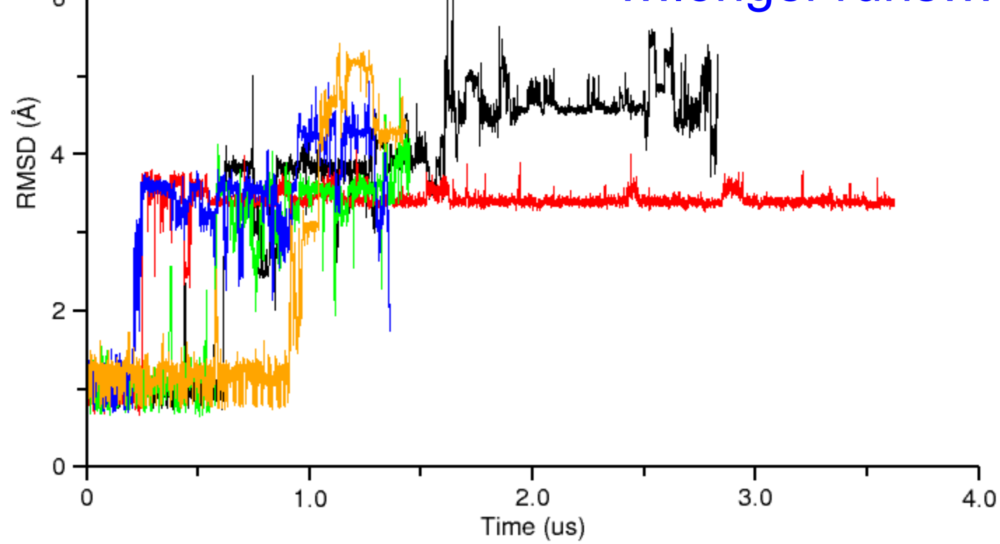
~197 ns/day!


DHFR (NVE) HMR 4fs 23,558 Atoms
2x K80 boards (4 GPUs)
423.69
1x K80 board (2 GPUs)
334.05
1/2x K80 board (1 GPU)
229.29
4X K40
489.68
2X K40
364.67
1X K40
266.07
2X K20
263.85
1X K20
196.99
1X K8
116.09
GTX-Titan-Z (2 GPU, full board)
356.48
GTX-Titan-Z (1 GPU, 1/2 board)
261.82
2X GTX Titan Black
383.32
1X GTX Titan Black
280.54
2X GTX 780
361.33
1X GTX 780
251.43
2X GTX 680
270.99
1X GTX 680
184.67
2X C2075
129.79
1X C2075
81.26
2xE5-2660v2 CPU (16 Cores)
30.21
0.00
100.00
200.00
300.00
400.00
500.00
Performance (ns/day)

Independent simulations of 2KOC “UUCG” tetraloop



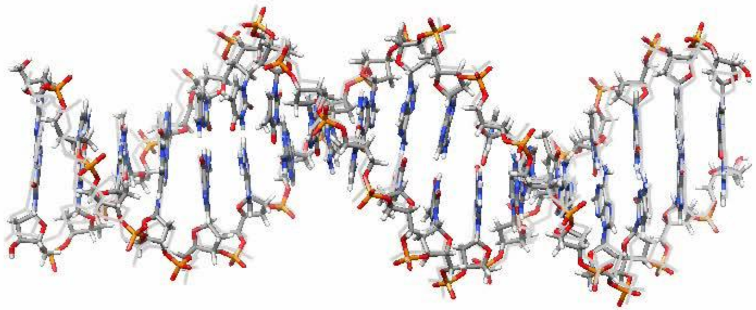
…longer runs…



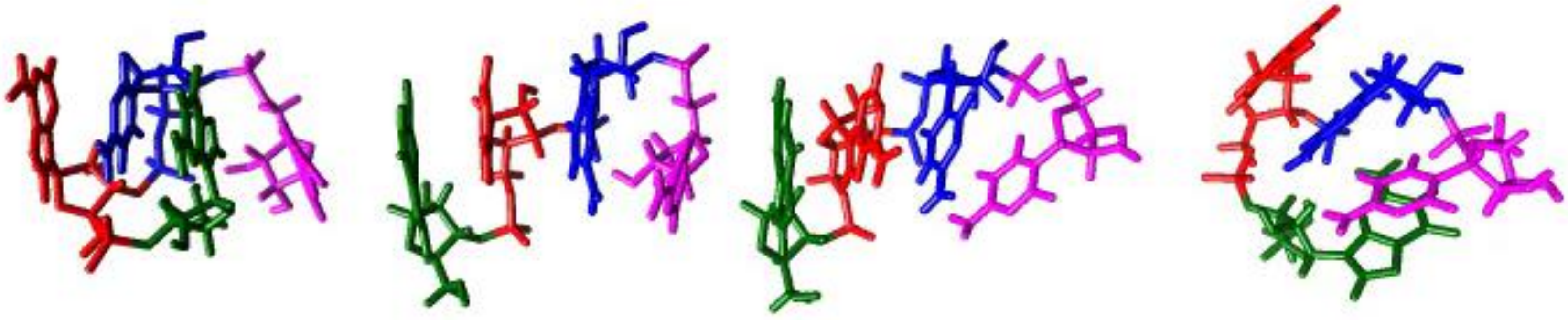
Limited sampling & too complex: Is there a simpler set of systems?

Today: two "long-time-to-develop" short stories…

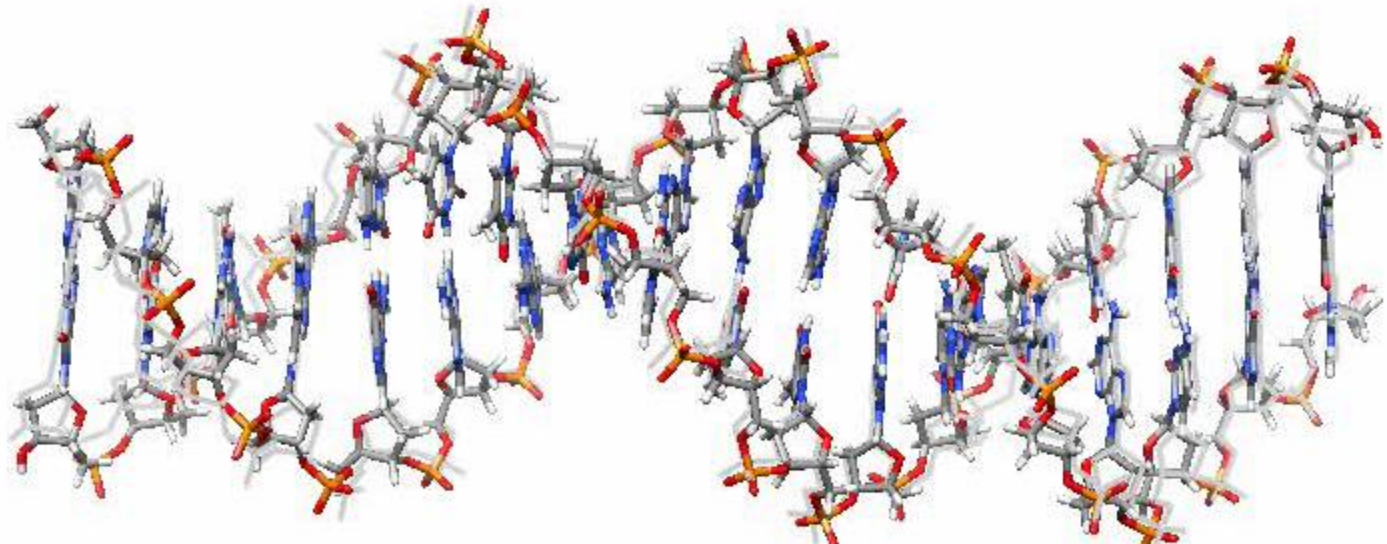
✓ **can we converge DNA duplex structure/dynamics?**



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



Anton run:



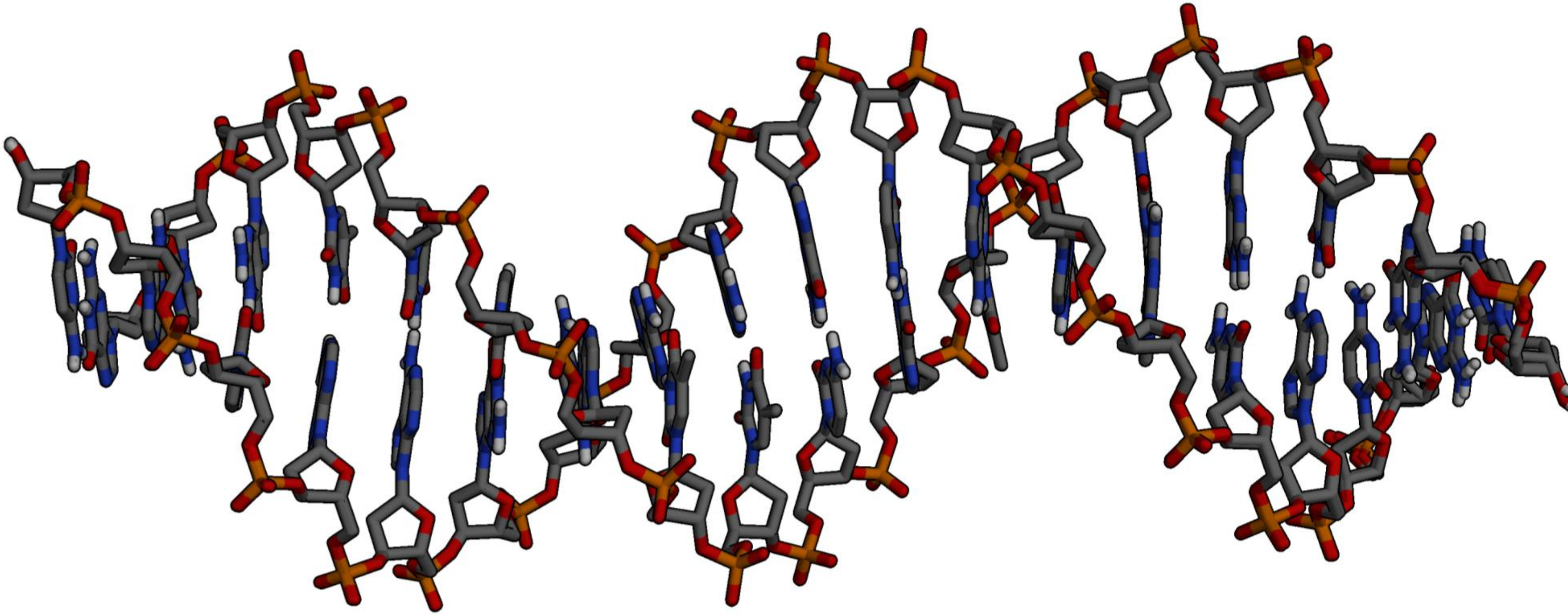
2 ns intervals, 10 ns running average, every 5th frame (~10 us).

~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 …

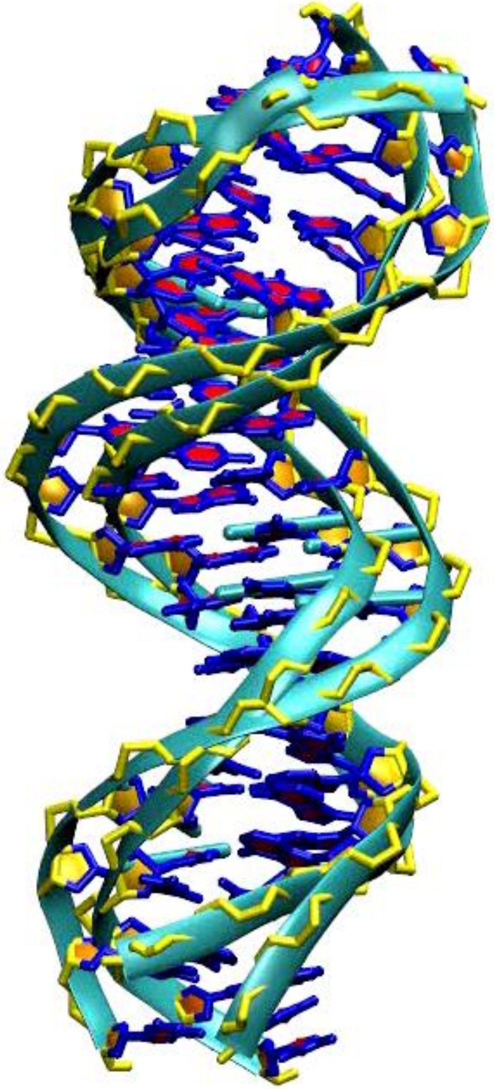
RMSd (0.028 Å) (0.049 Å) (0.076 Å) (0.160 Å)



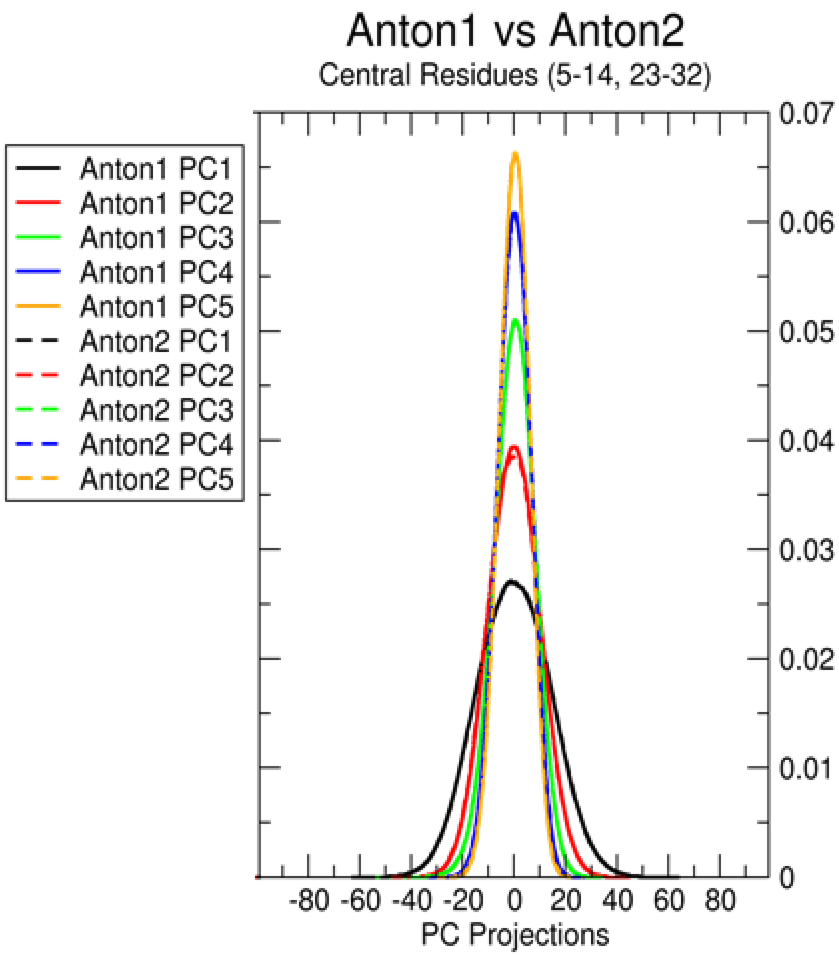
…this cannot be right, can it?

(breathing, bending, twisting, …)

Test for convergence within and between simulations: Dynamics
Principal components (or major modes of motion)

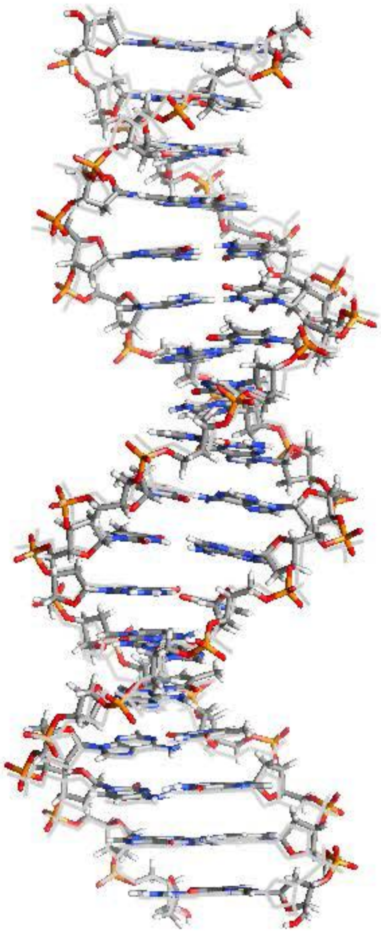


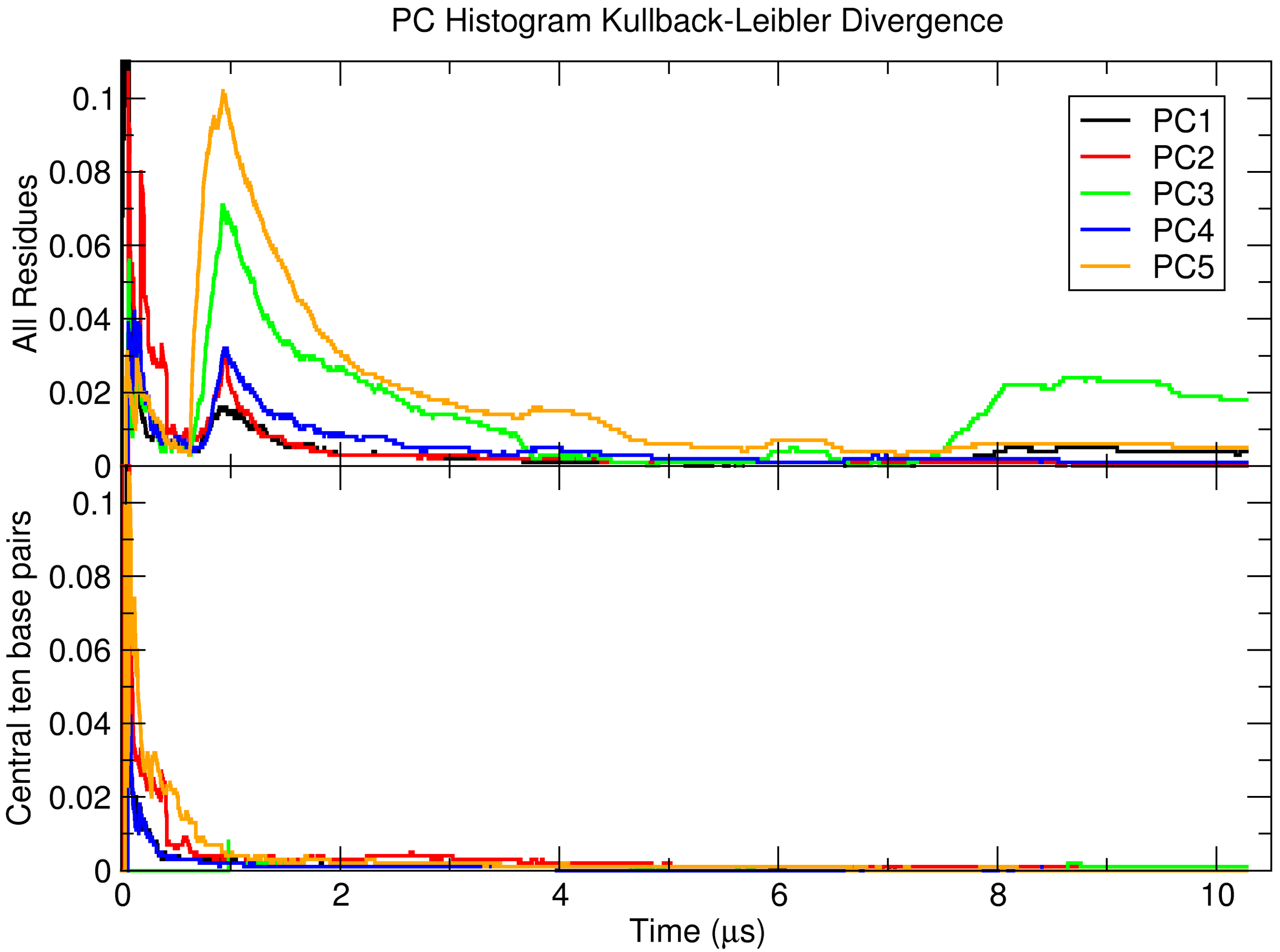
Visualization of the first two (dominant) modes of motion



Overlap of modes from independent simulations (internal helix)

Test for convergence within and between simulations: How long does it take to converge the PC's?





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

all tetraloops

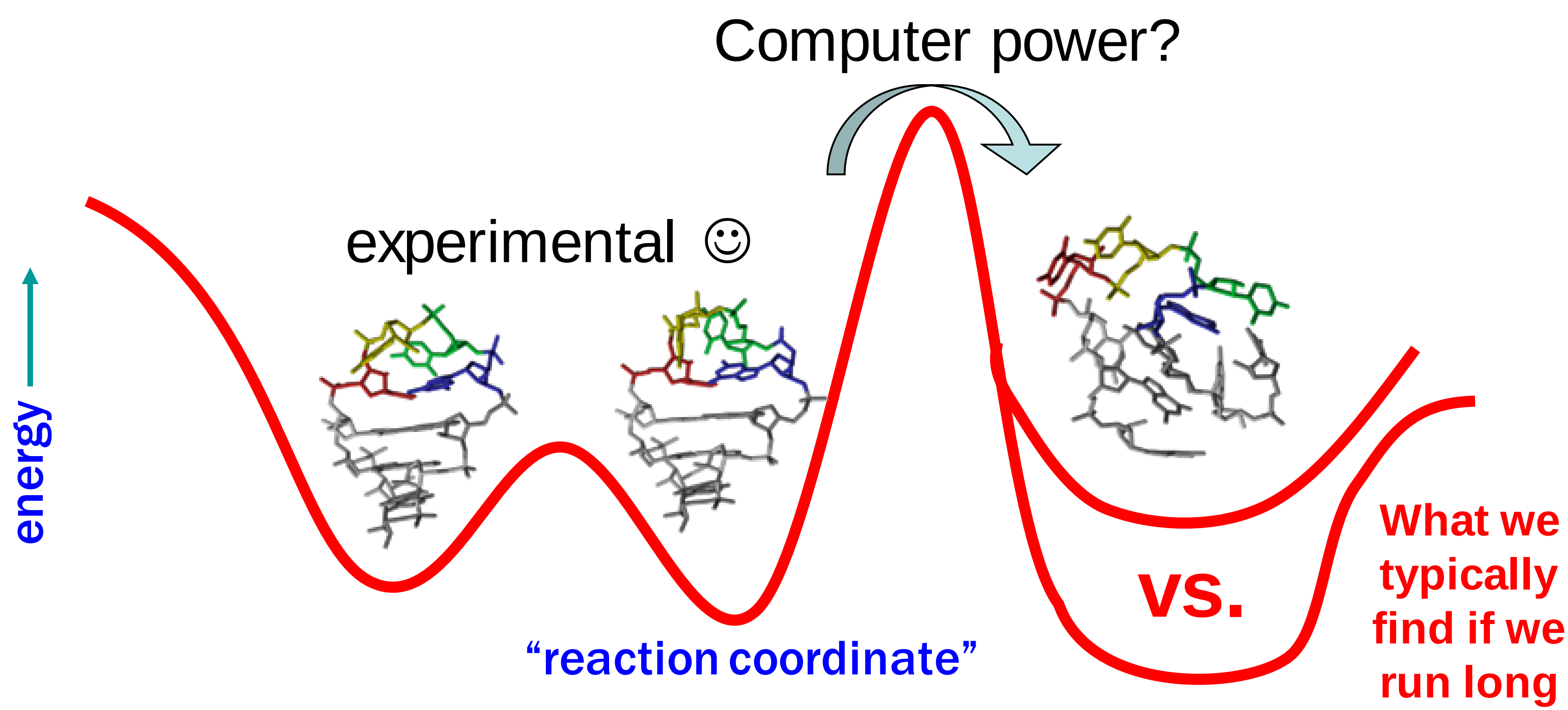
NMR structures of DNA & RNA

crystal simulations

RNA motifs

quadruplexes

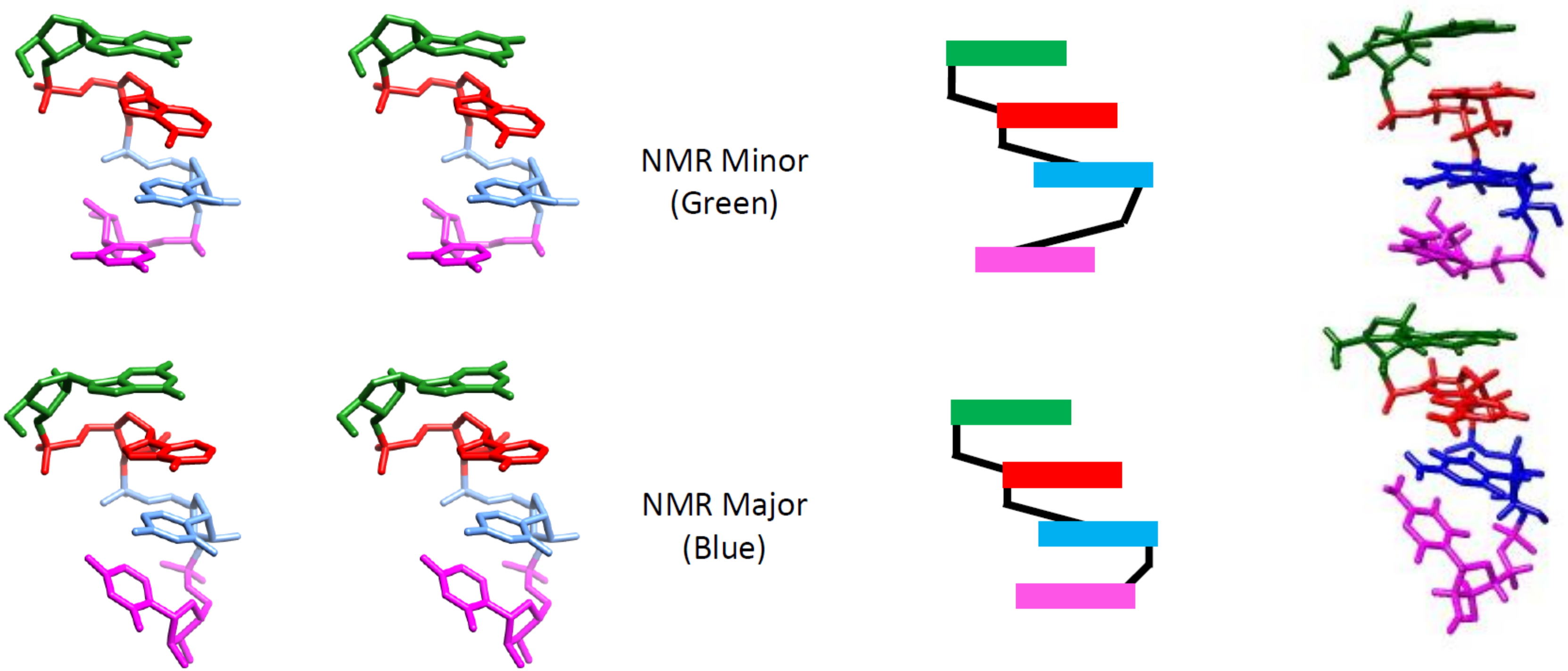
RNA-drug interactions



…a system where we can get complete sampling

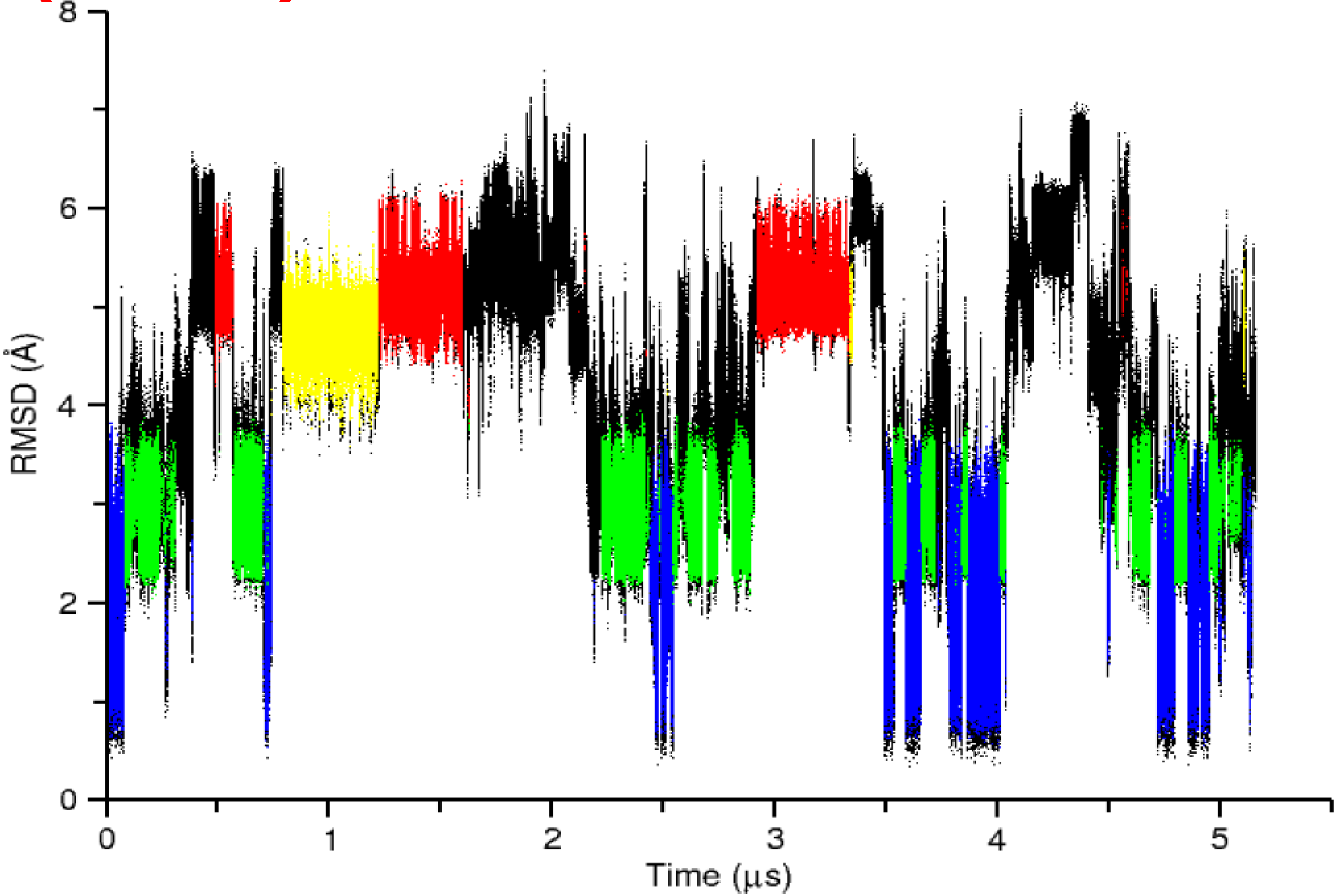
r(GACC) tetranucleotide

[Turner / Yildirim]

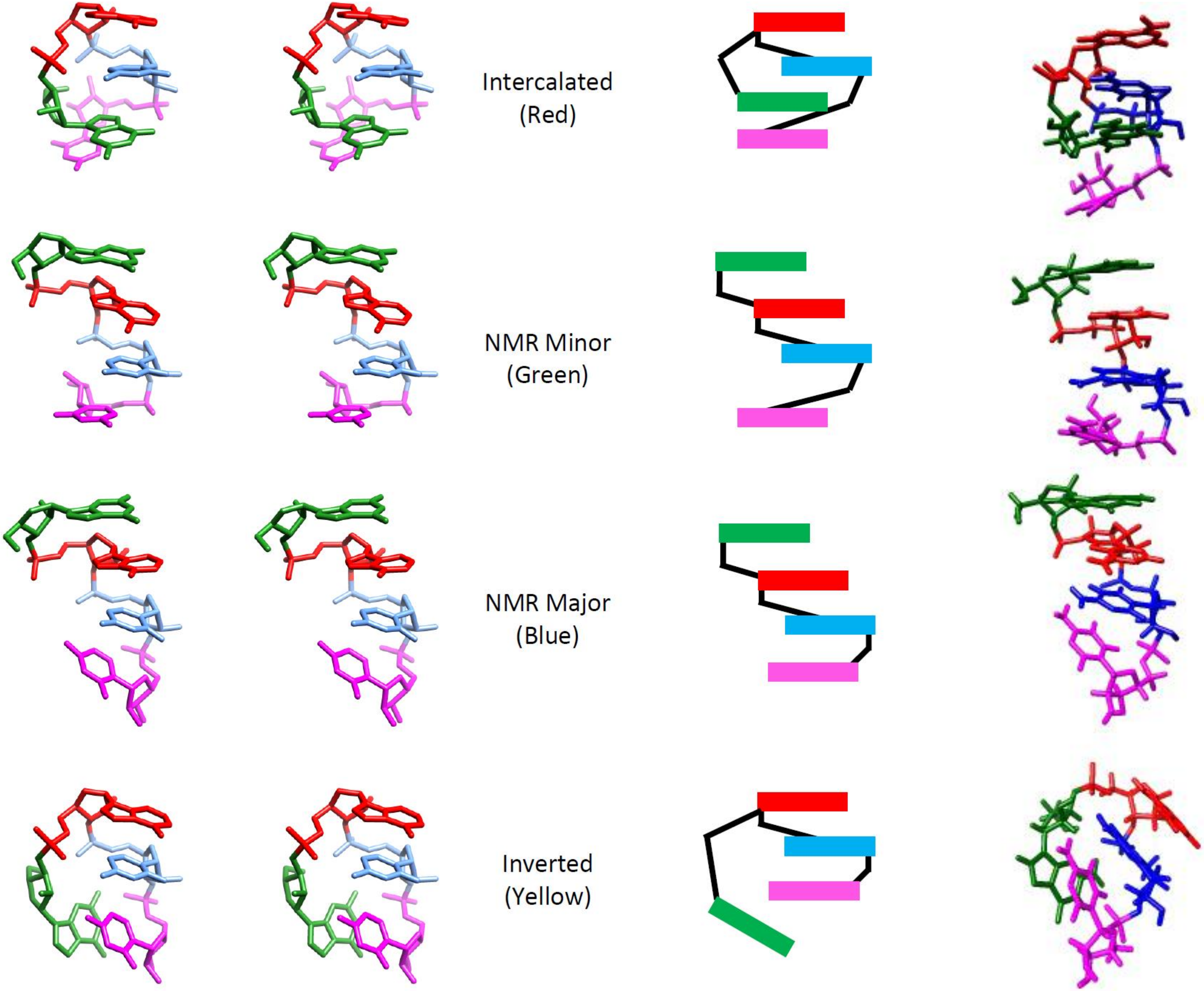


NMR suggests two dominant conformations…
…compare to MD simulations in explicit solvent

r(GACC) tetranucleotide: AMBER ff12



< explicit solvent time-contiguous MD >


Intercalated
(Red)
NMR Minor
(Green)
NMR Major
(Blue)
Inverted
(Yellow)

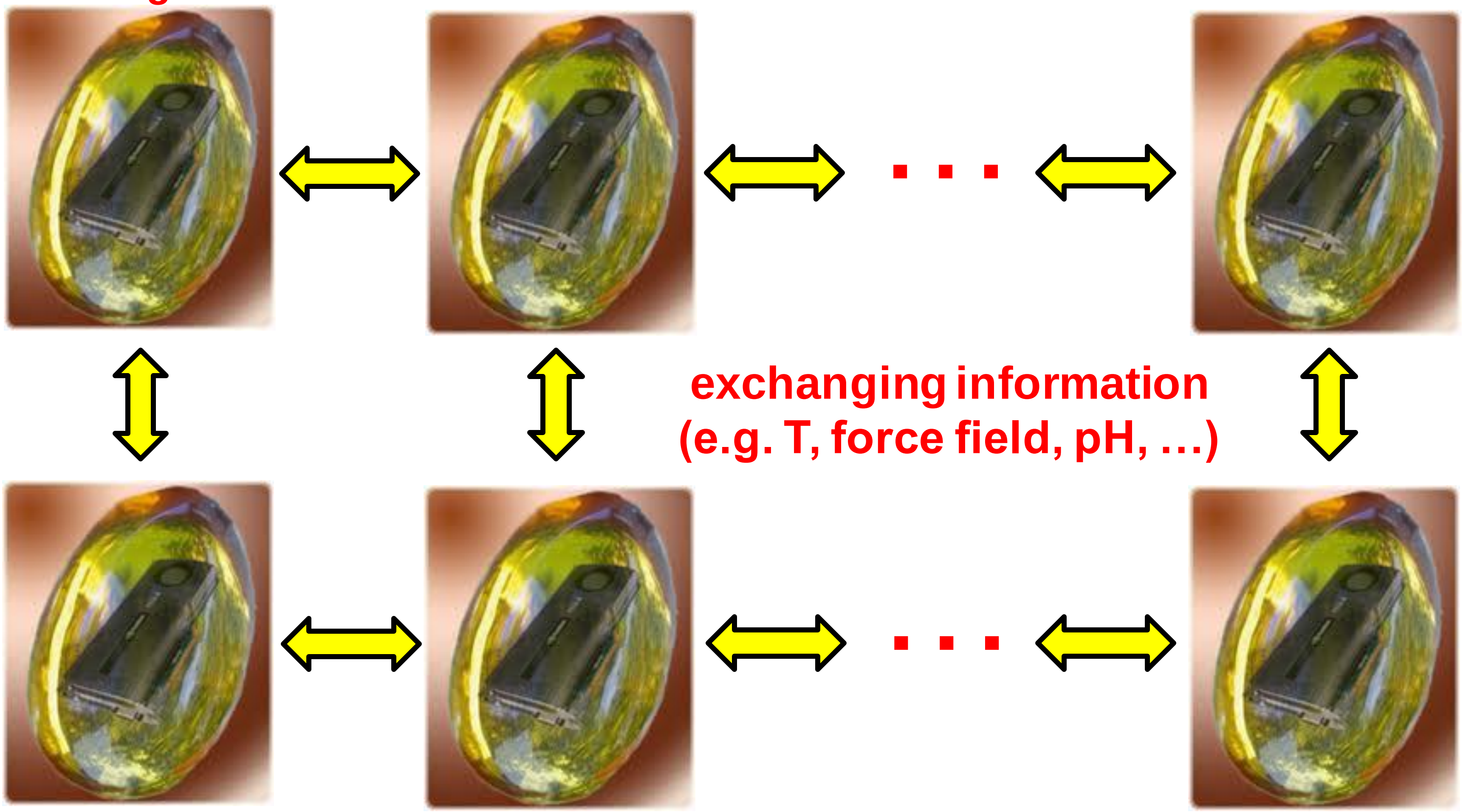
…still need more sampling!

(enablers)

- strong GPU performance of AMBER/PMEM
- good replica exchange functionality
- access to Keeneland, Stampede, Blue Wate

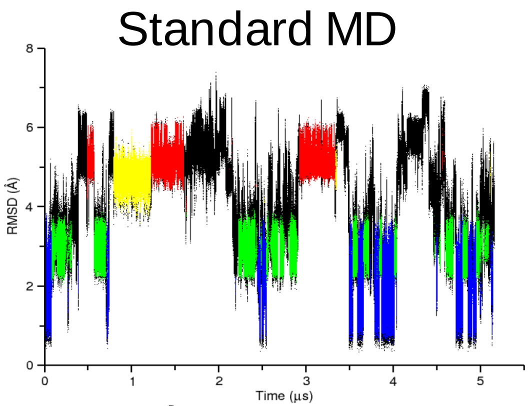
Blue Waters PRAC: The main goals are to hierarchically and tightly couple a series of optimized molecular dynamics engines to fully map out the conformational, energetic and chemical landscape of RNA.

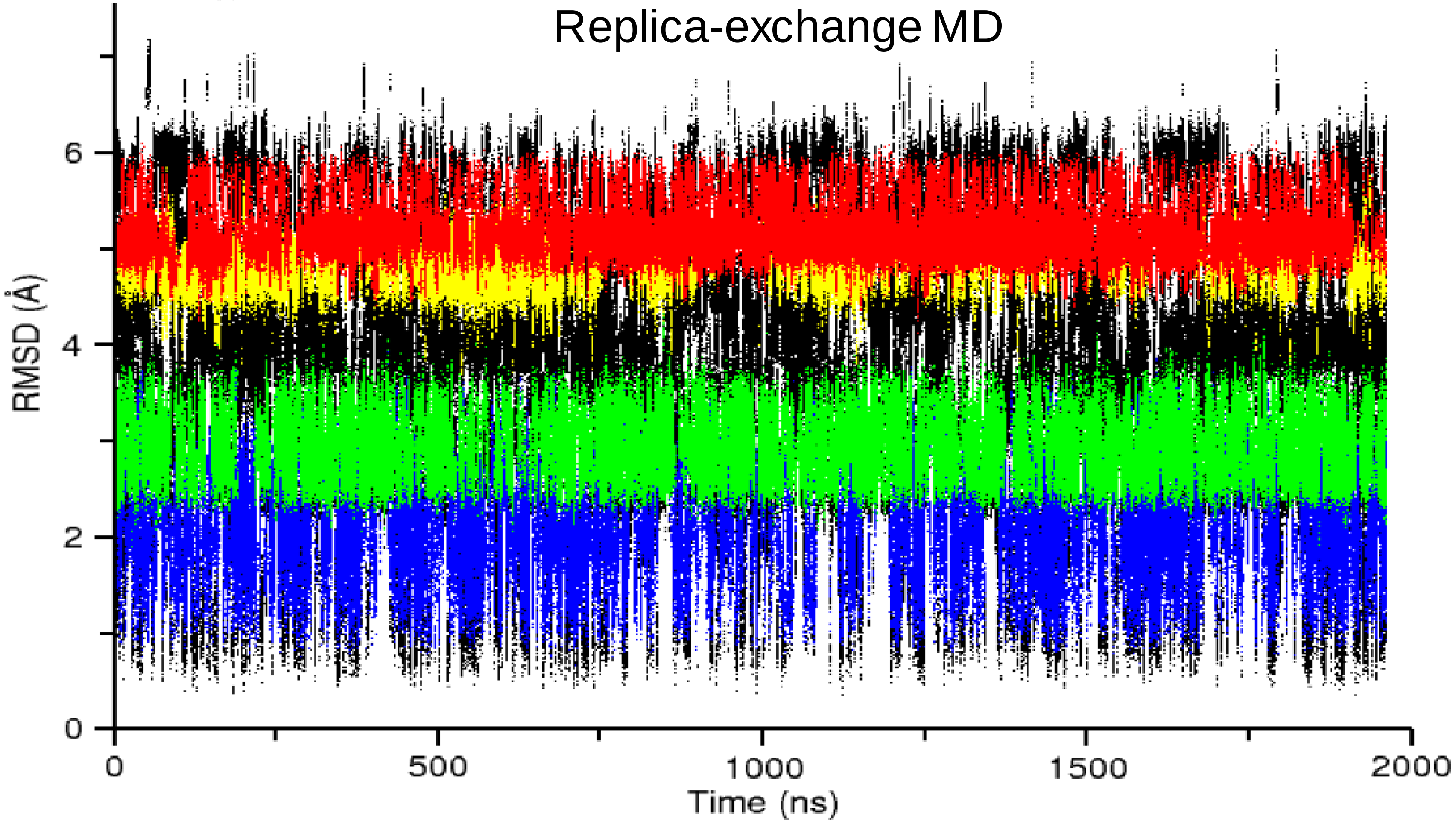
independent || MD engines



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

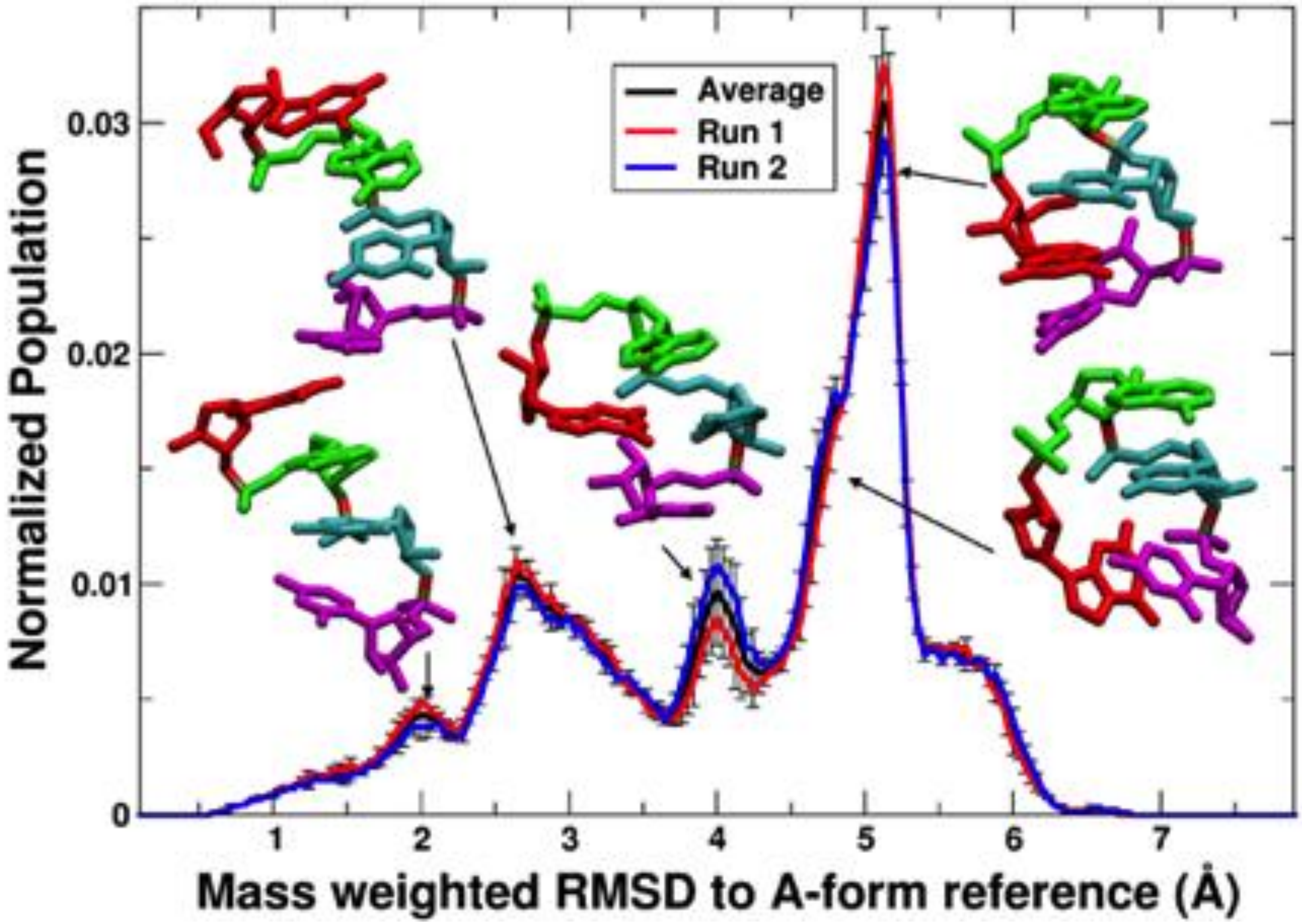
r(GACC) tetranucleotide



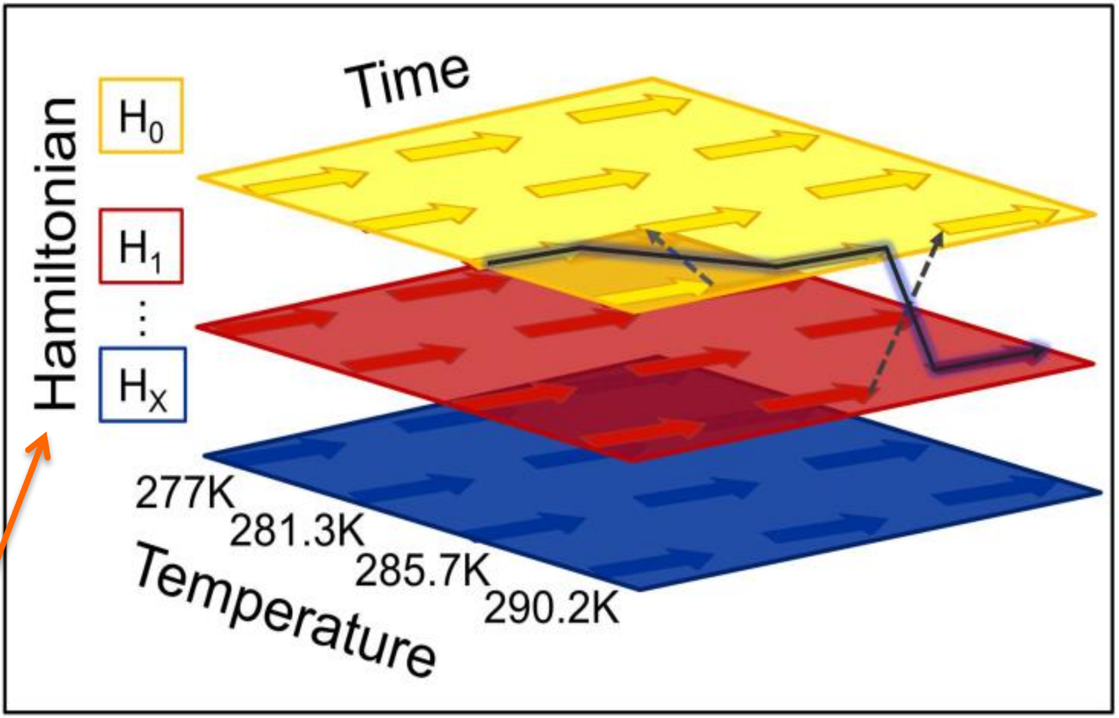


RMSD distribution profiles: Distance from A-form reference

(aka each peak shows population certain distance from the reference)



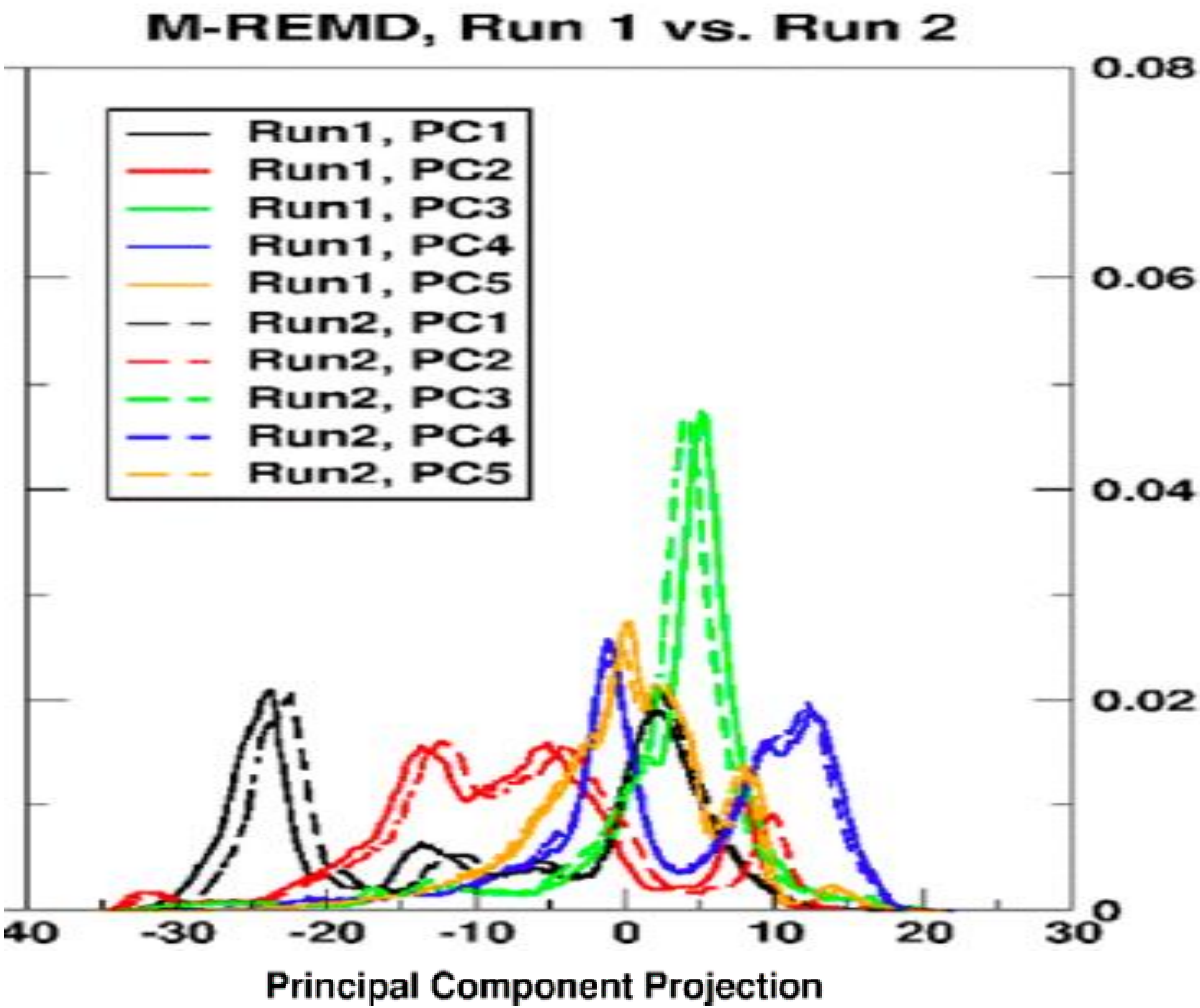
multi-D REMD – Bergonzo / Roe, Roitberg / Swa

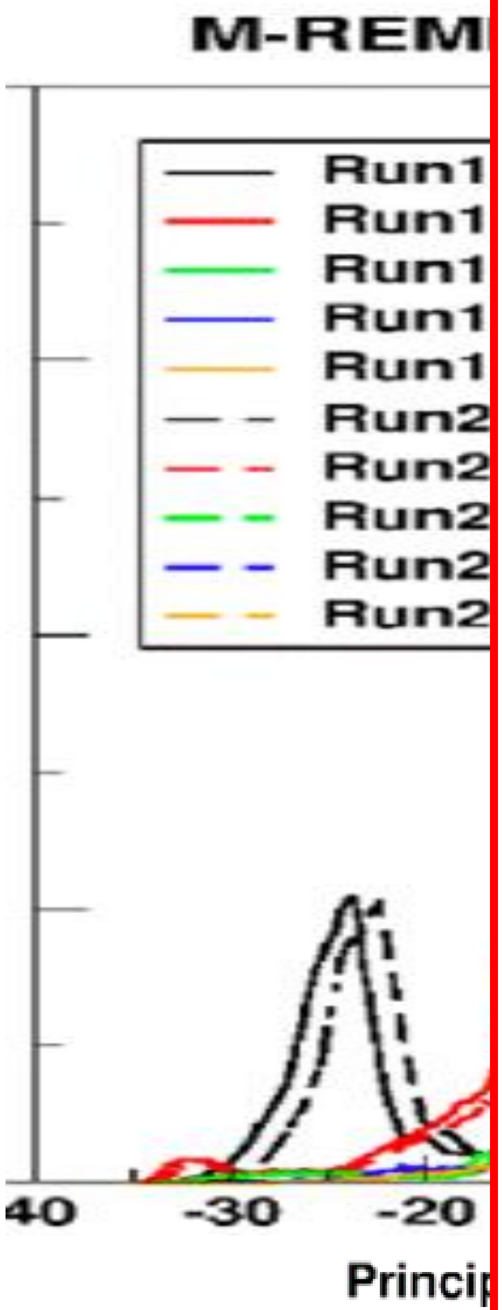


Change in "energy representation"

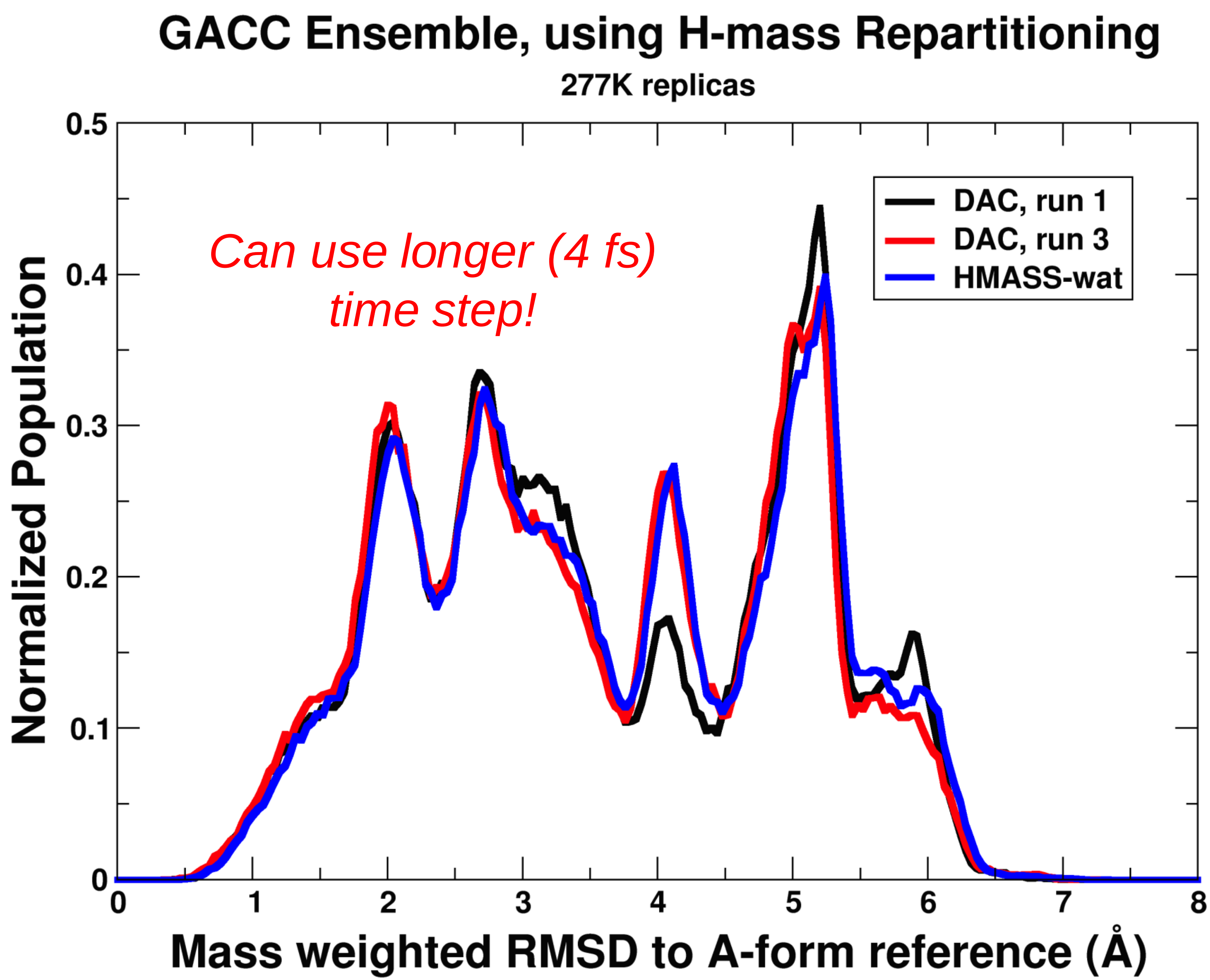
- **pH**
- **restraints, umbrella potentials, …**
- **force field / parameter sets**
- **biasing potentials (aMD)**

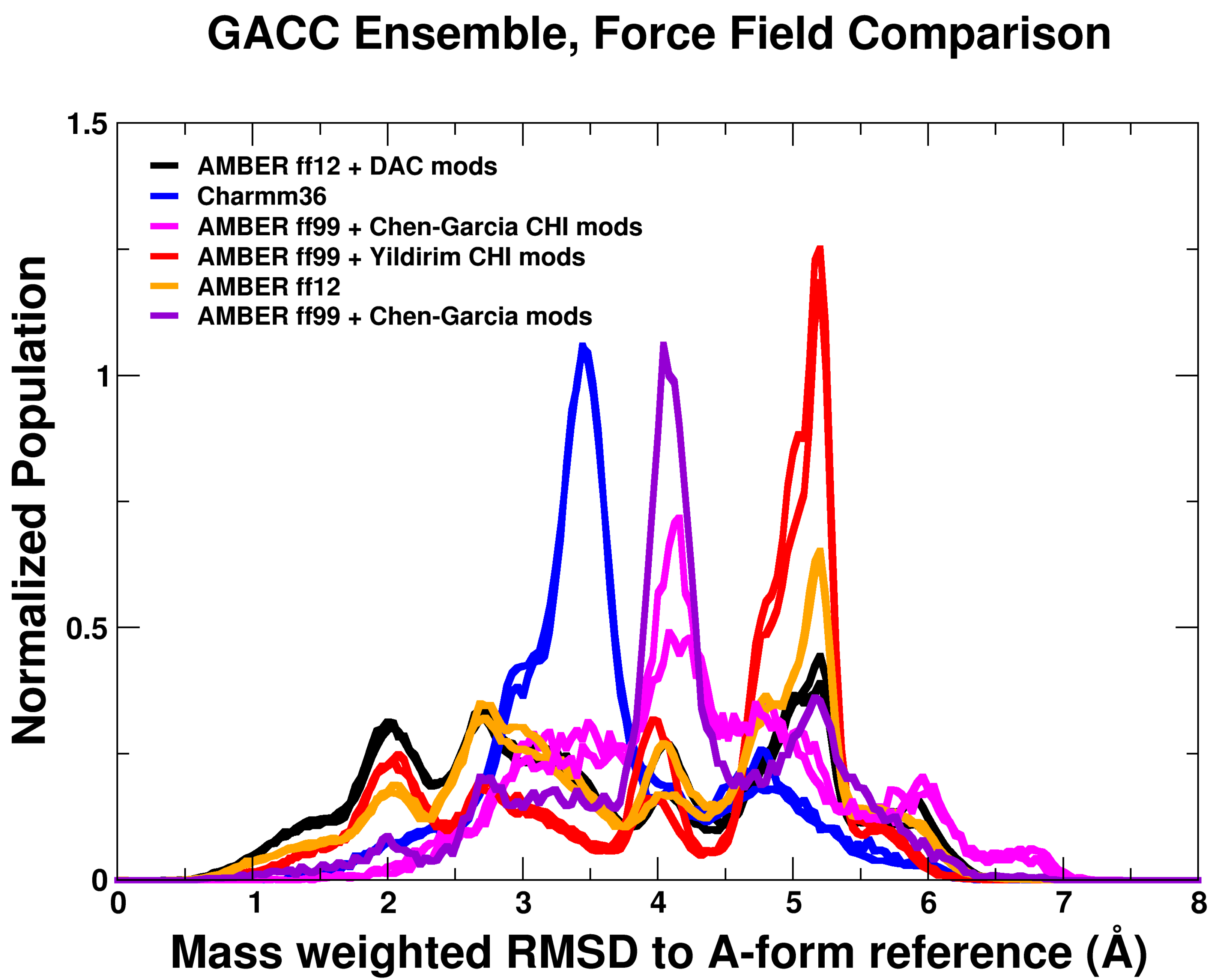
Fukunishi, H., Wanatabe, O., and Takada, S., J. Chem. Phys. 2002.
Sugita, Y., Kitao, A., and Y. Okamoto, J. Chem. Phys. 2000.


M-REMD, Run 1 vs. Run 2
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
-40
-30
-20
-10
0
10
20
30
Principal Component Projection



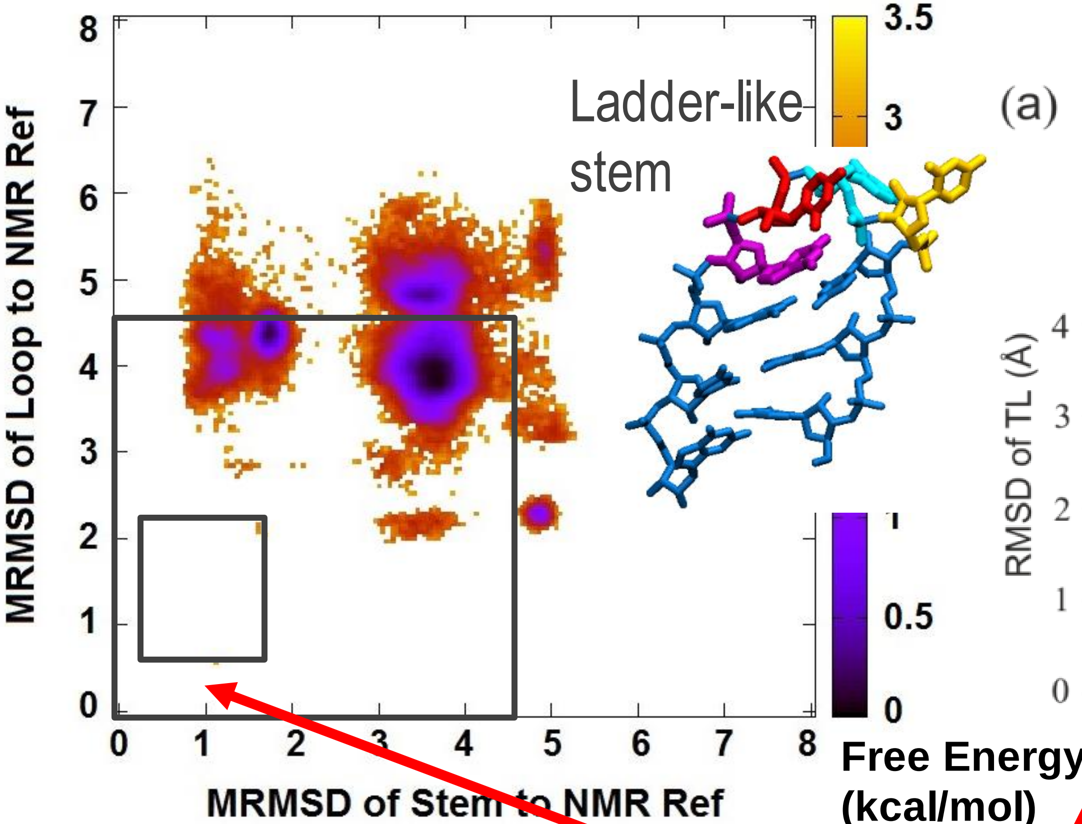

```
# Read in both trajectories
#
trajin traj.run1.nc
trajin traj.run2.nc
# RMS-fit to first frame
#
rms first :1-4&!@H=
# Create an average structure
#
average gaccAvg.rst7 ncrestart
# Save coordinates as 'crd1'
#
createcrd crd1
run
# Fit to average structure
#
reference gaccAvg.rst7.1 [avg]
# RMS-fit 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
```


GACC Ensemble, using H-mass Repartitioning
277K replicas
Can use longer (4 fs) time step!
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

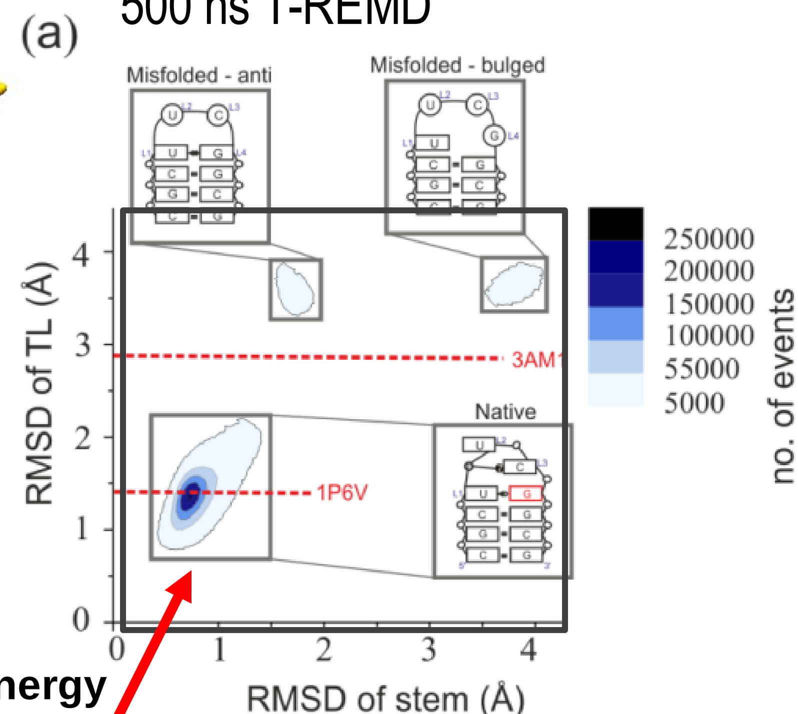

GACC Ensemble, Force Field Comparison
AMBER ff12 + DAC mods
Charmm36
AMBER ff99 + Chen-Garcia CHI mods
AMBER ff99 + Yildirim CHI mods
AMBER ff12
AMBER ff99 + Chen-Garcia mods
Normalized Population
1.5
1
0.5
0
0
1
2
3
4
5
6
7
8
Mass weighted RMSD to A-form reference (Å)

...more complete sampling alters results

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



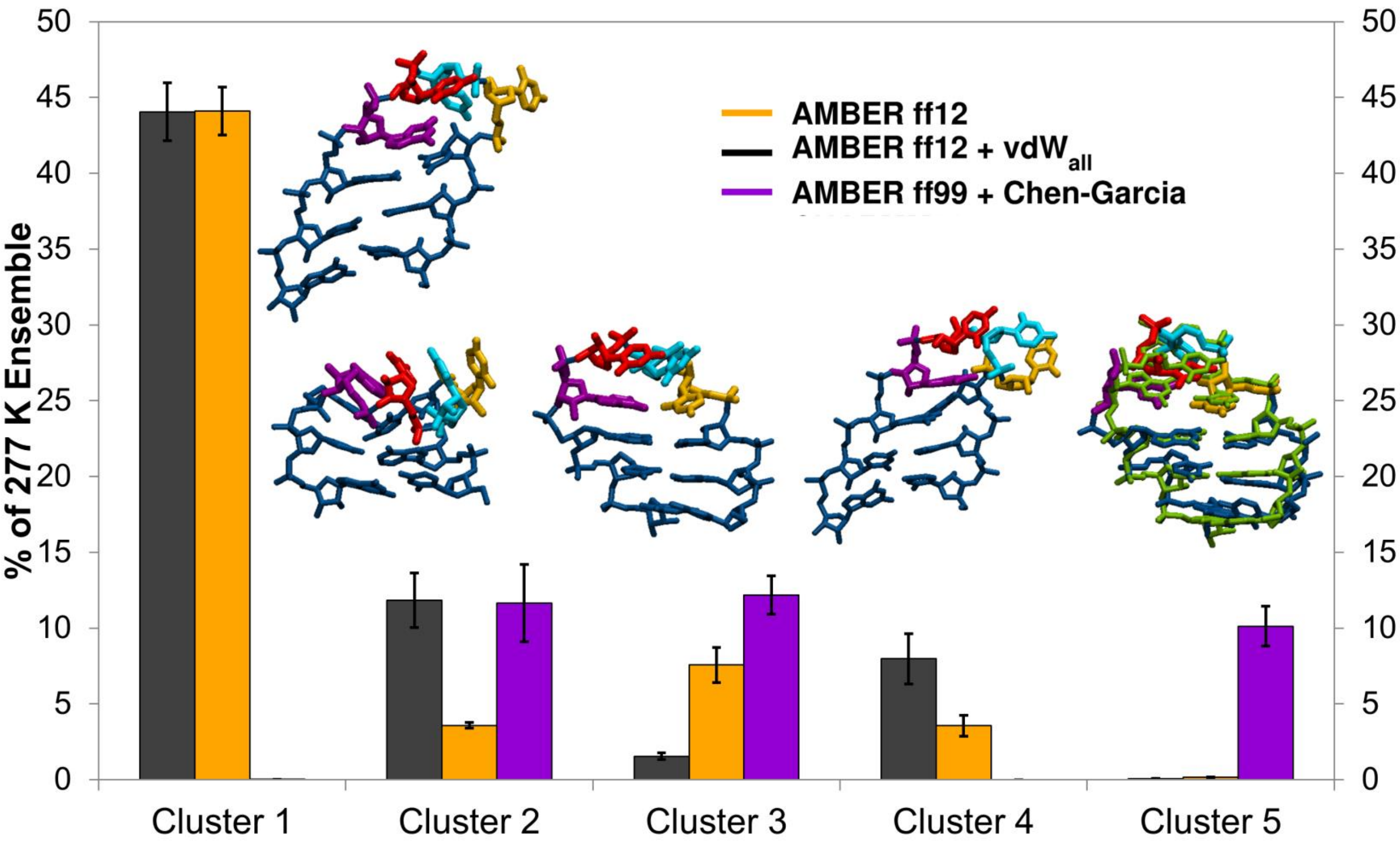
Kührová et al. 2013 JCTC
500 ns T-REMD

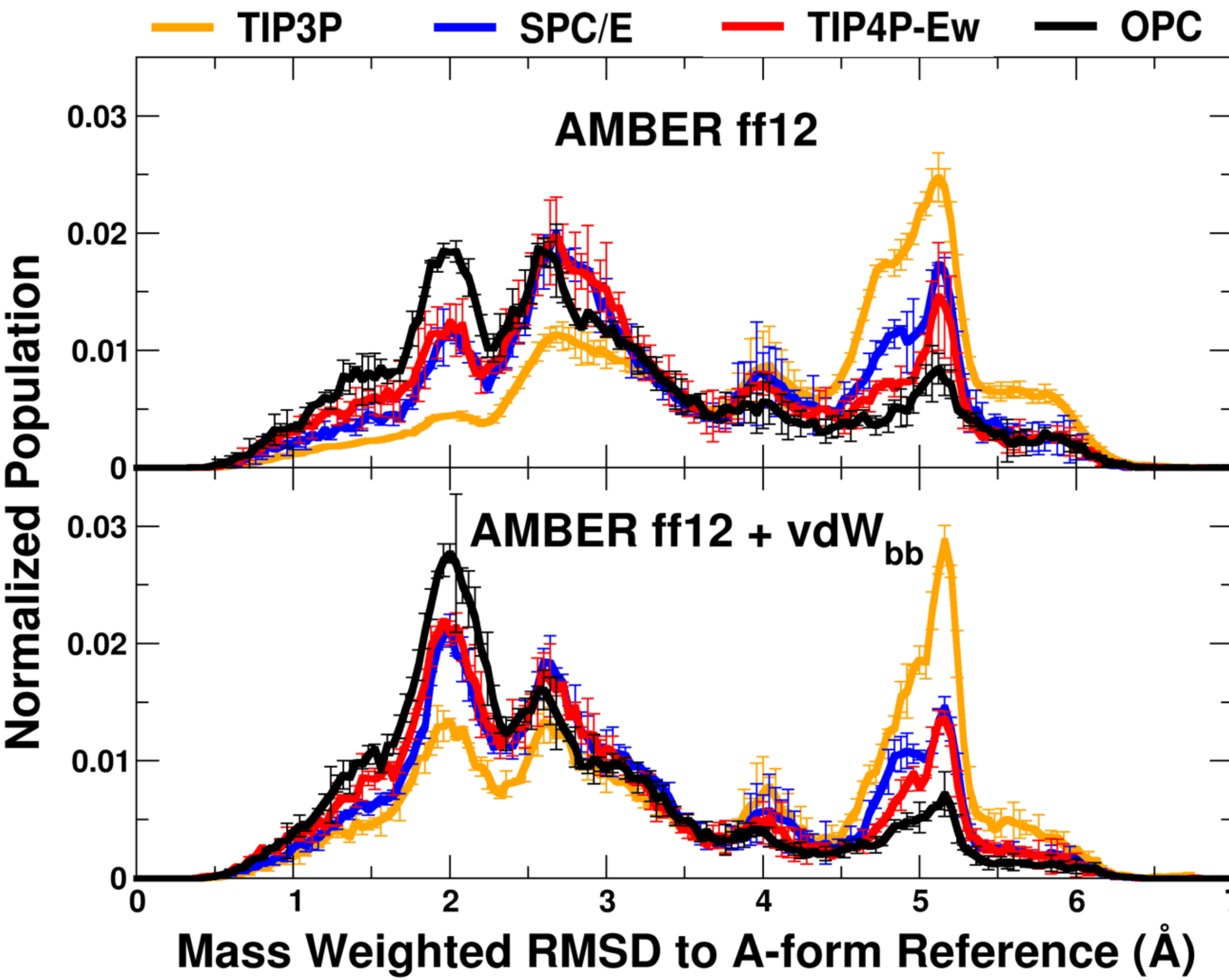


Native

ff99 Chen-Garcia shifts the population

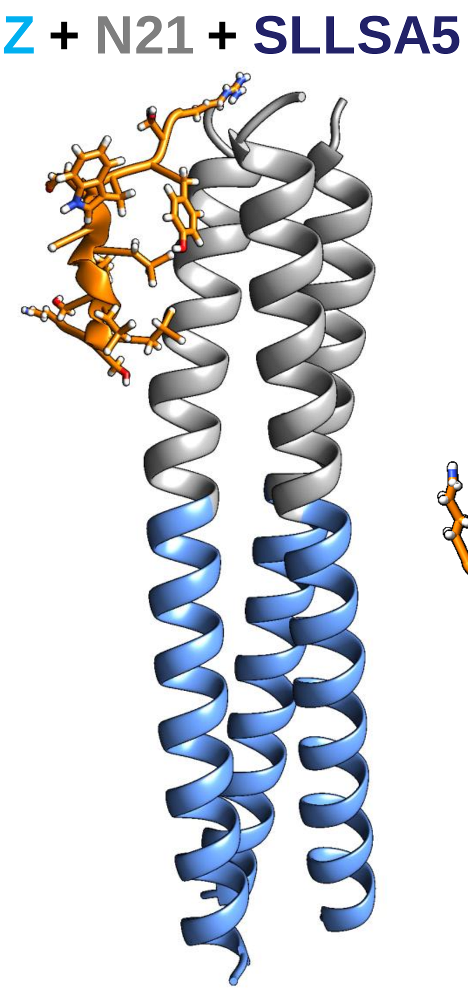
- Folded UUCG tetraloop structure is sampled
- Iso-energetic structures



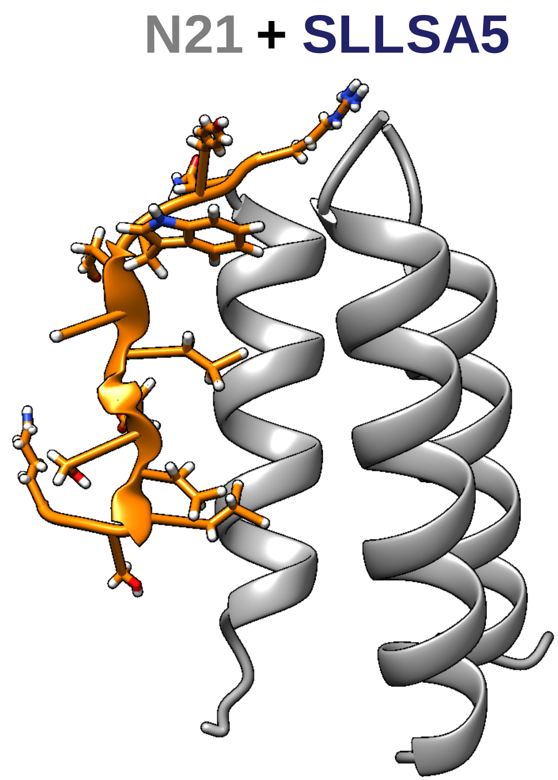


r(GACC): We now get correct 3:1 population of experimental structures with anomalous structures < 5%

FUTURE: Ebola membrane fusion inhibitor peptide design



Avg
Structure
(150 ns sim)

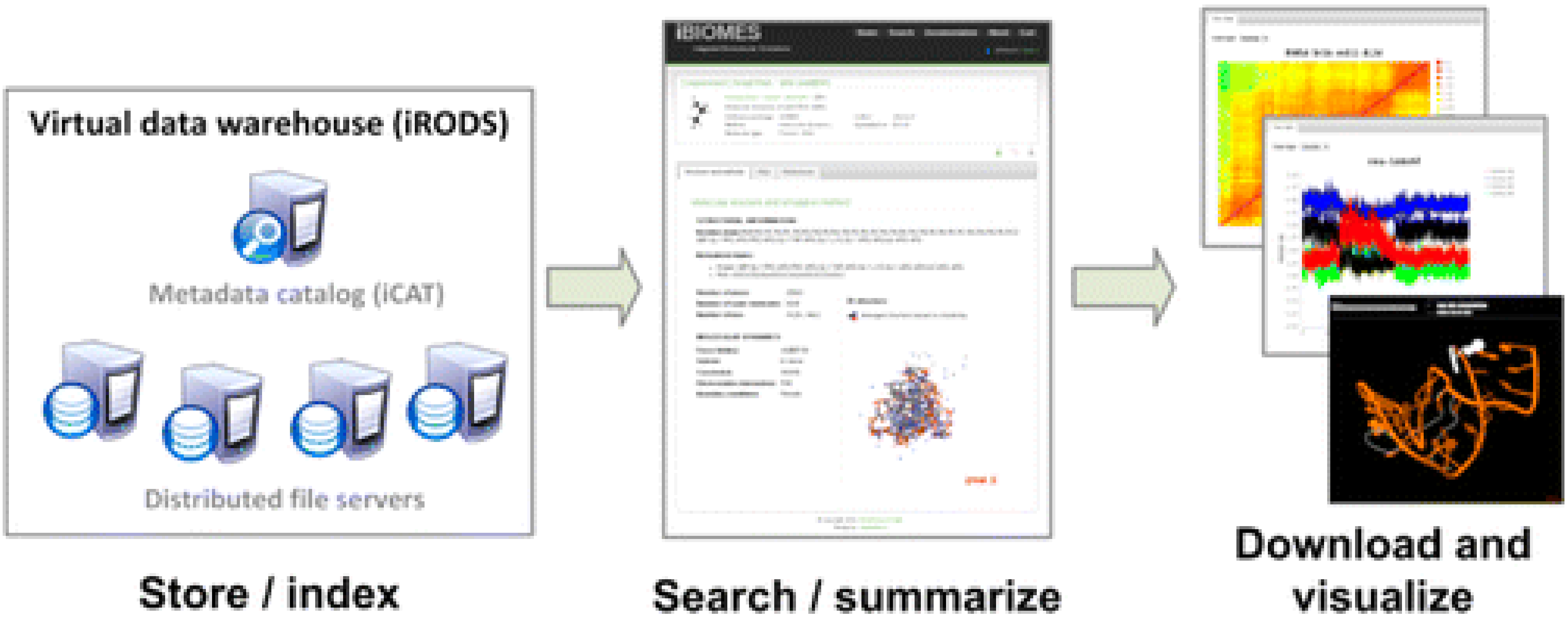


Avg
Structure
(220 ns sim)

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

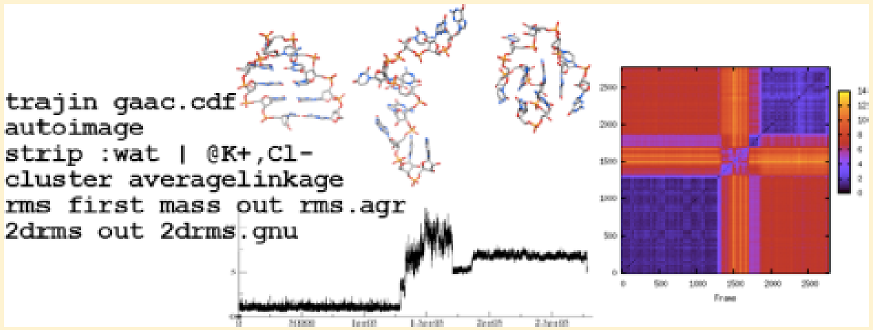
pubs.acs.org/JCTC

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?