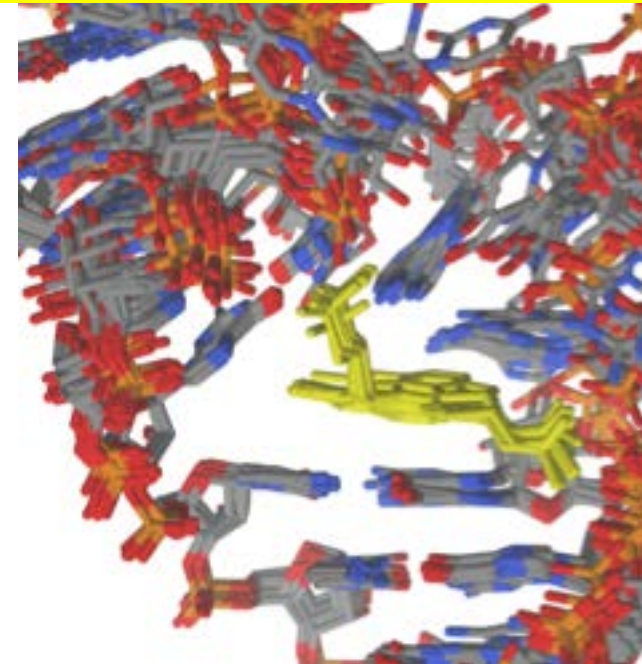
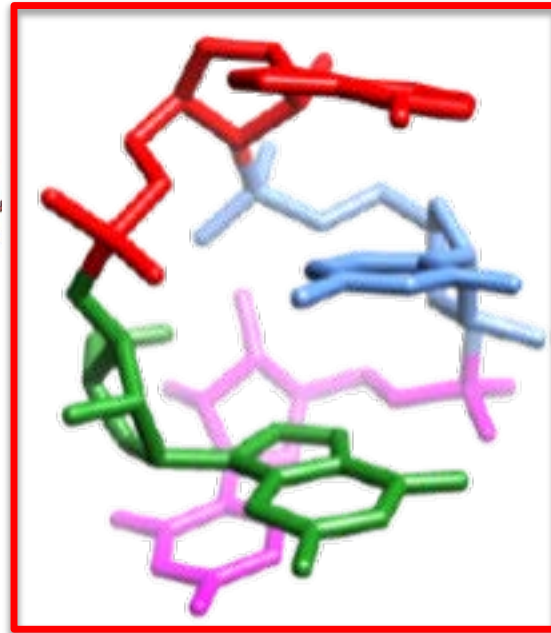
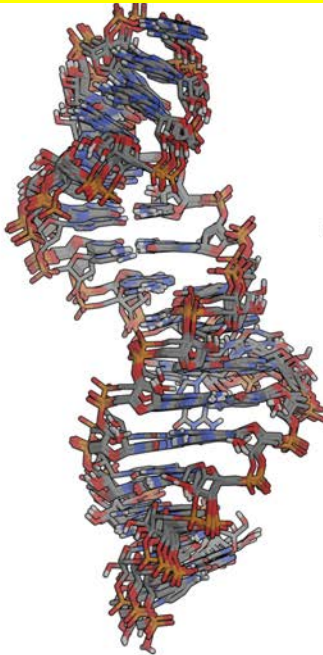
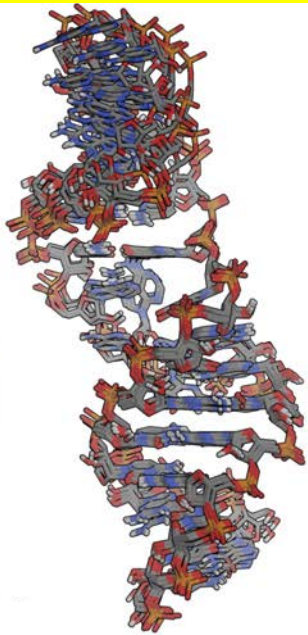


# Convergence, reproducibility and accuracy in the simulation of conformational ensembles of nucleic acids: Surprise!



**Thomas E. Cheatham III**

**tec3@utah.edu**

**Professor, Department of Medicinal Chemistry, College of Pharmacy**

**Director, Research Computing and the  
Center for High Performance Computing**

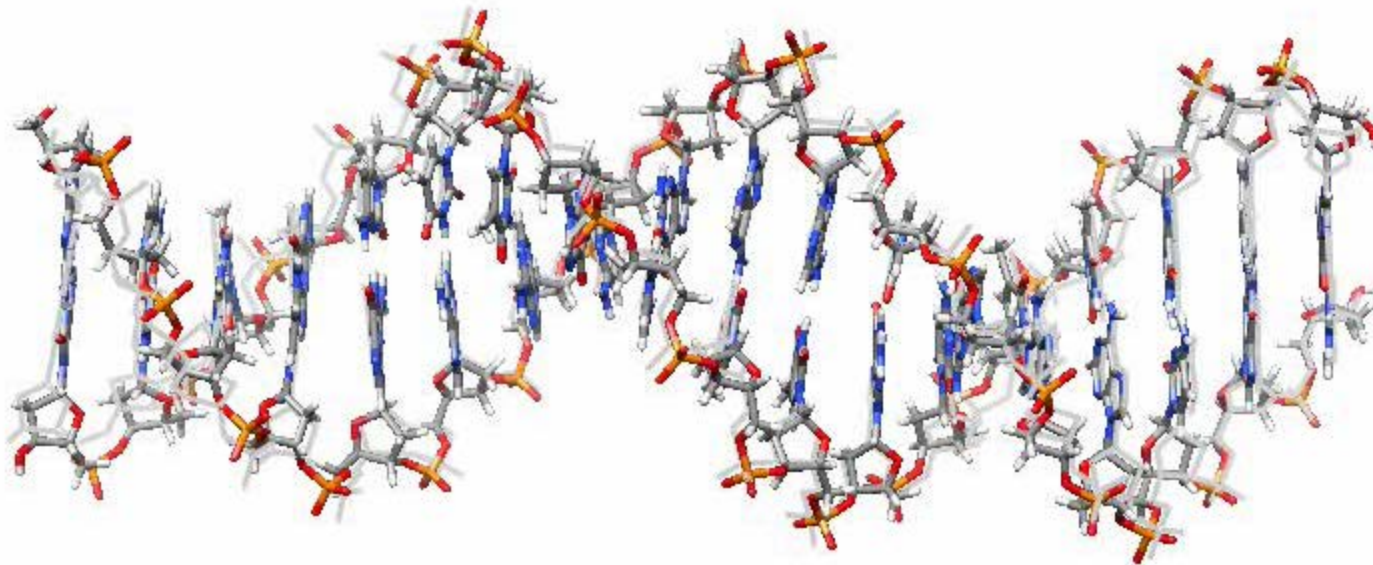
**University Information Technology**

**University of Utah**

# biomolecular simulation

...structure, dynamics, interactions,  $\Delta G$ ,  
sampling, force fields

---



**AMBER ff, MD on Anton1@PSC** – data at 2 ns intervals,  
10 ns running average, every 5<sup>th</sup> frame (~10  $\mu$ s of MD shown).

*reproducibility, convergence, agreement with experiment, new insight*



**What does this research require?**

**...computing support...**

**physical and people resources**

**locally & nationally**



~1-2M core hours / year  
~500 TB RAID disk

XSEDE

Extreme Science and Engineering  
Discovery Environment

~10M core hours / year  
*Award: MCA01S027*  
XSEDE SAB / UAC

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

~12M *node* hours / year  
Multiple PB of data  
*Award: PRAC ACI-1515572*  
*Ebola RAPID ACI-1521728*  
Blue Waters SETAC

**What is needed to properly set-up, run, assess and validate simulations of nucleic acids aimed at elucidating the “converged” conformational ensemble?**

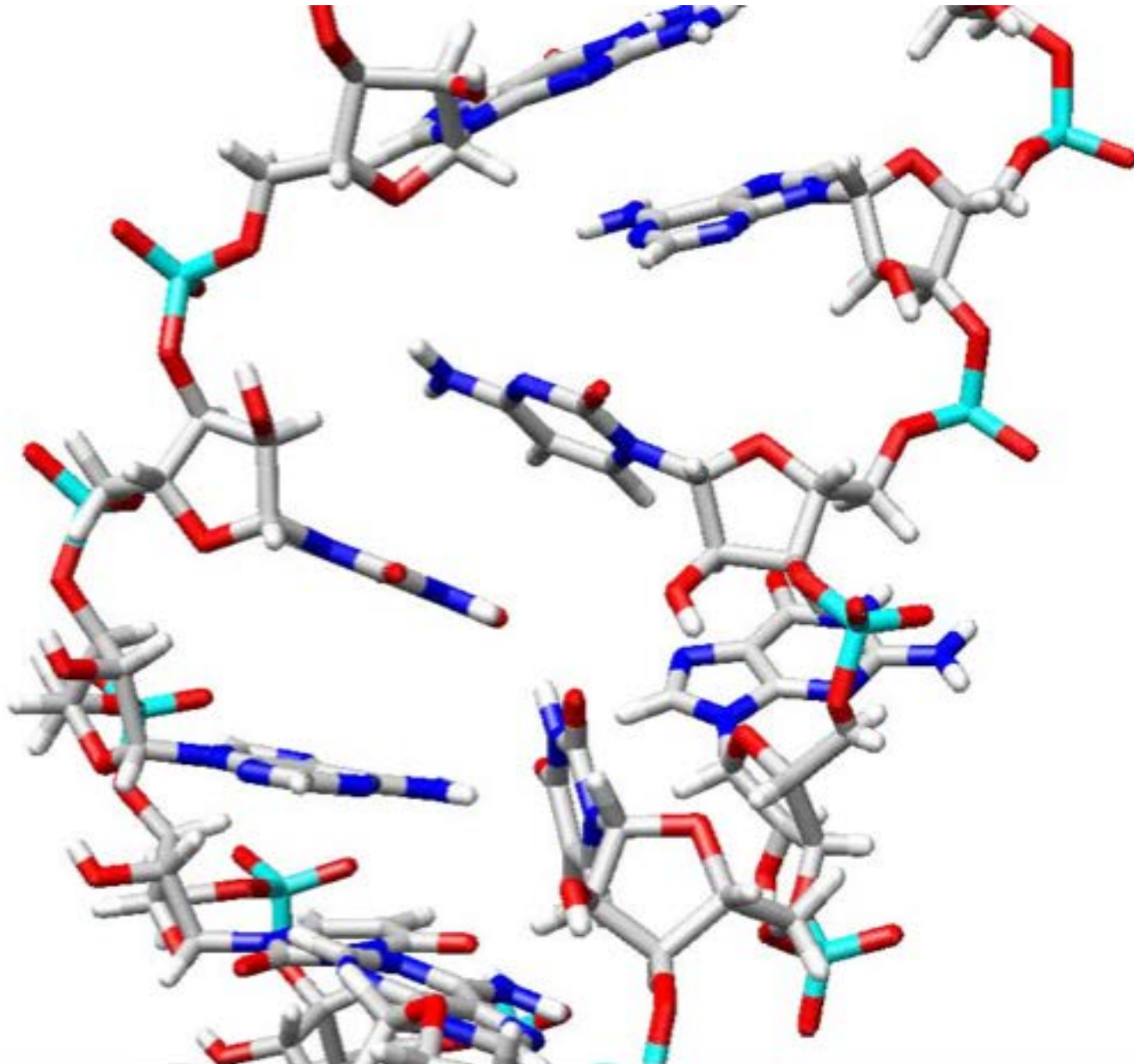
**What is needed to properly set-up, run, assess and validate simulations of nucleic acids aimed at elucidating the “converged” conformational ensemble?**

**Initial conditions:**

- **starting structures, set-up (force fields, ions, water), equilibration?**

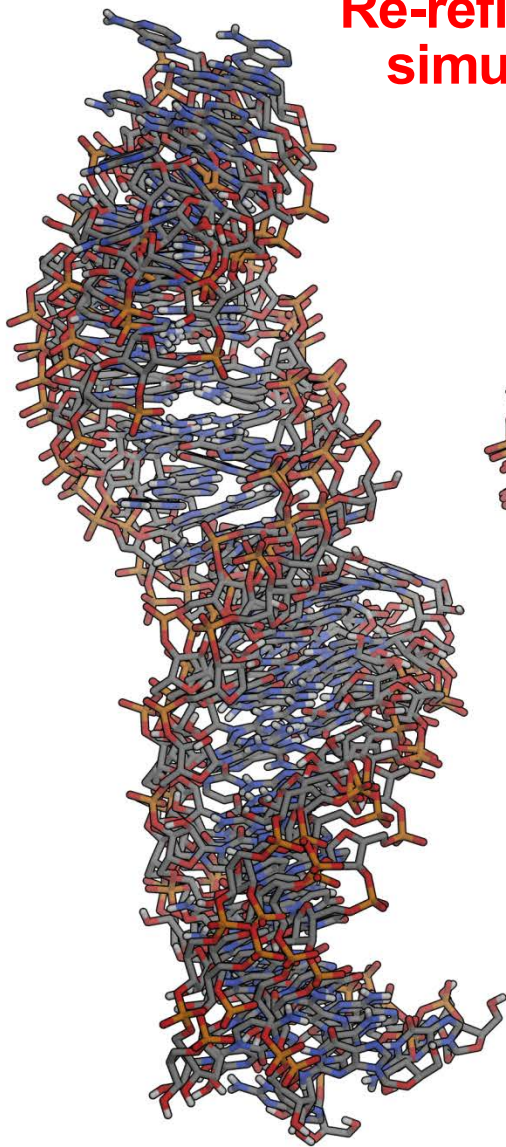
# MD simulation of a published group II intron ribozyme piece

PDB: 1R2P (~50 ns, smoothed): starting structure = NMR, ending structure ☹️

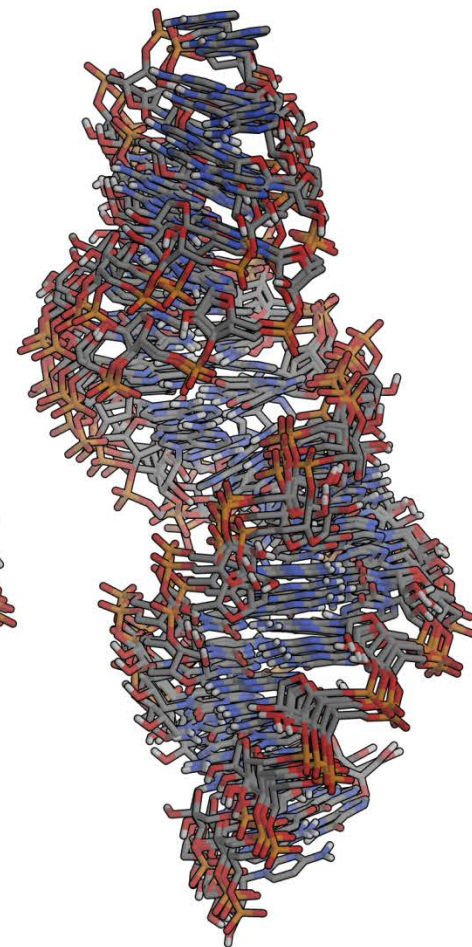
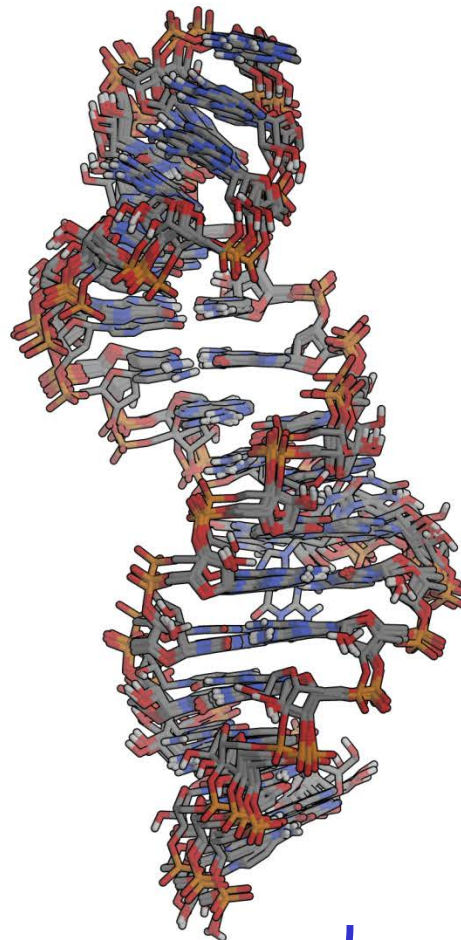
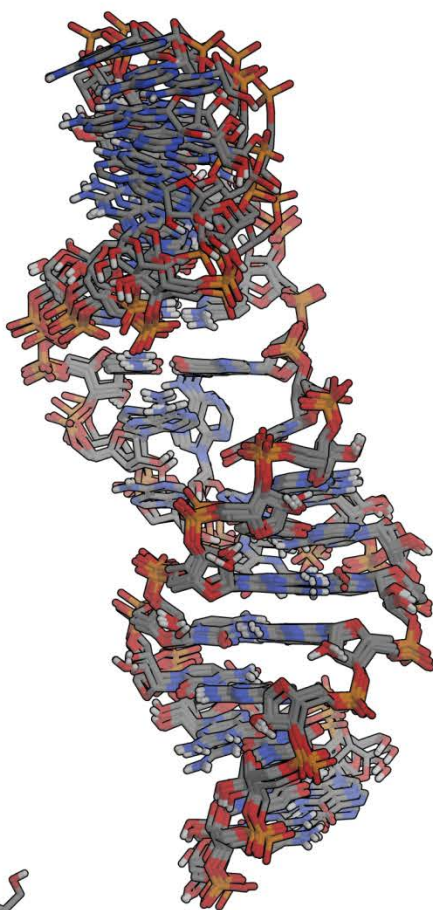


**Re-refinement of NMR helpful before MD simulation (on older RNA structures)**

N. Henricksen  
D.R. Davis



**NMR: 1R2P**




**NMR: 2F88**

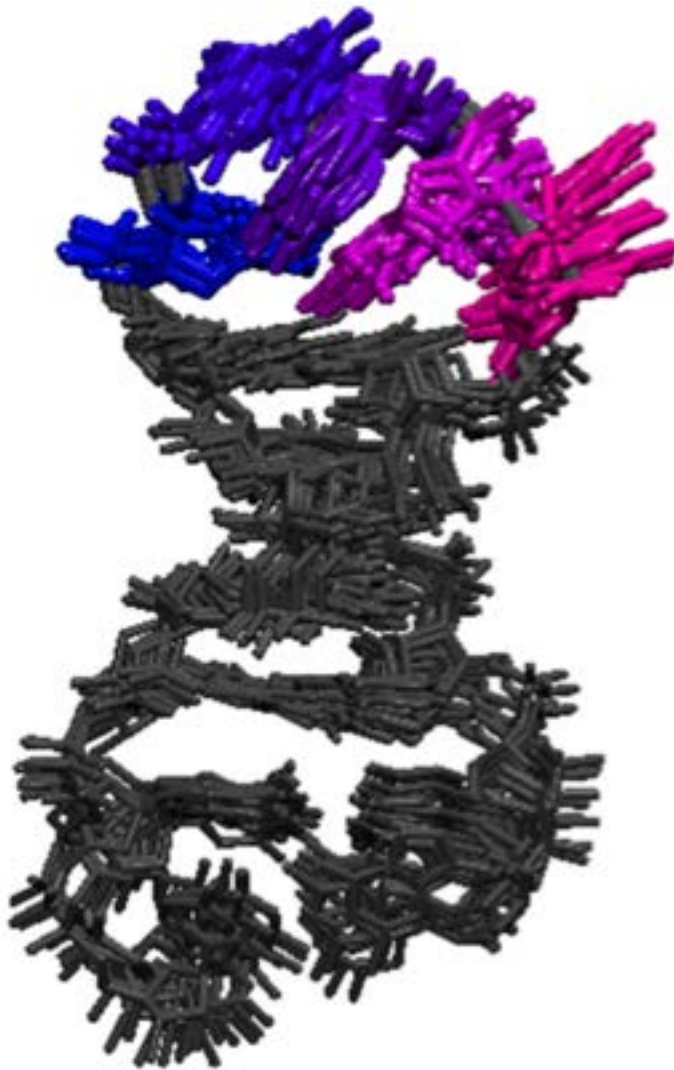


**simulated w/ restraints,  
modern force field,  
explicit solvent**

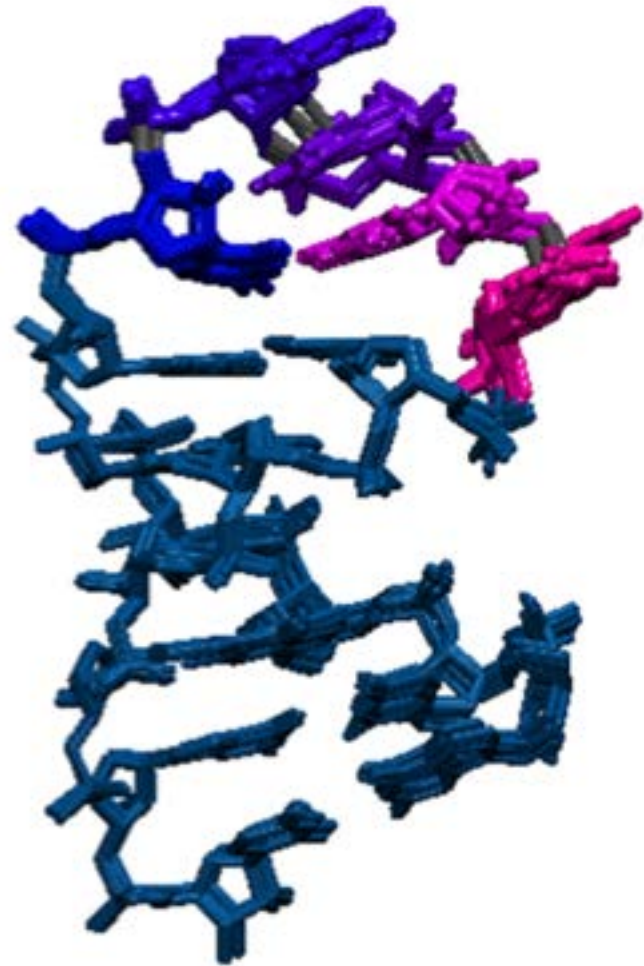


decoy: 1TBK  1YN2  $\pm$  Mg<sup>2+</sup>


-Mg<sup>2+</sup> deviates from NMR structure: re-refine...



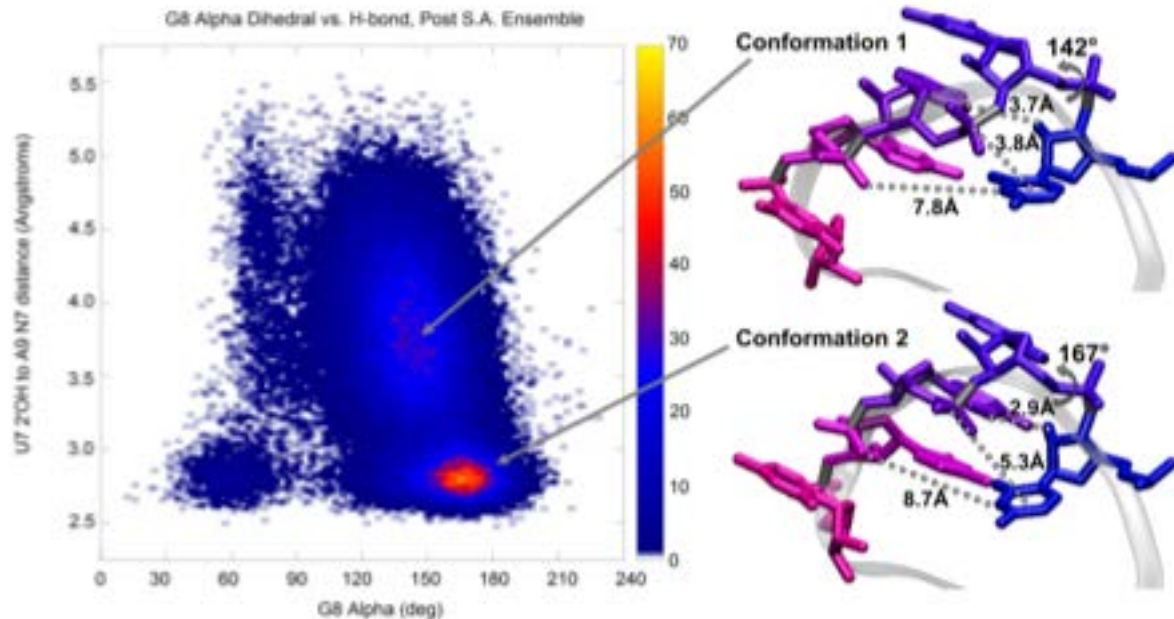
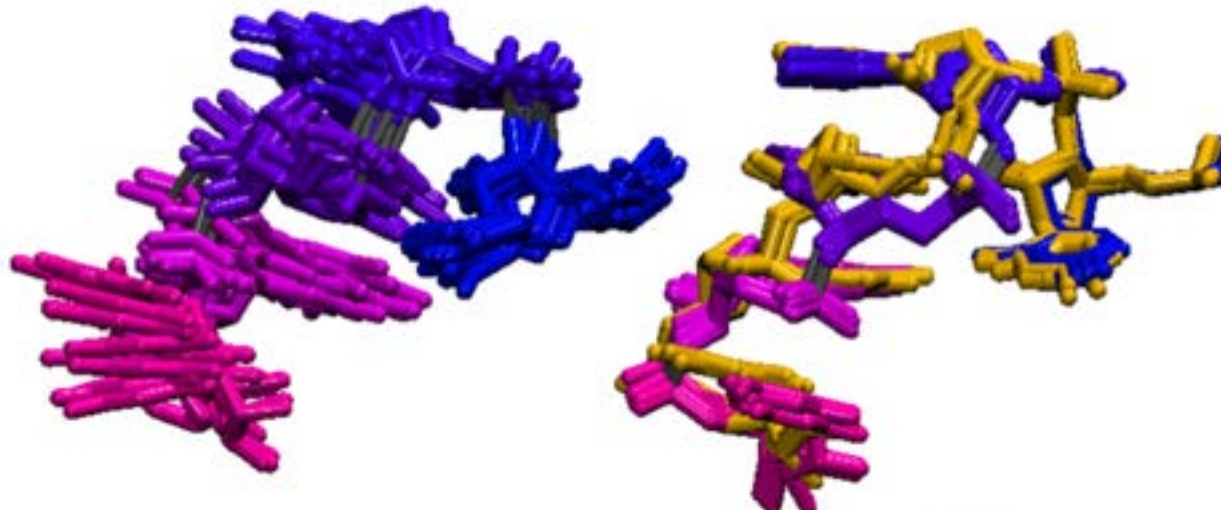
original NMR



re-refined NMR

decoy: 1TBK  1YN2  $\pm$  Mg<sup>2+</sup>

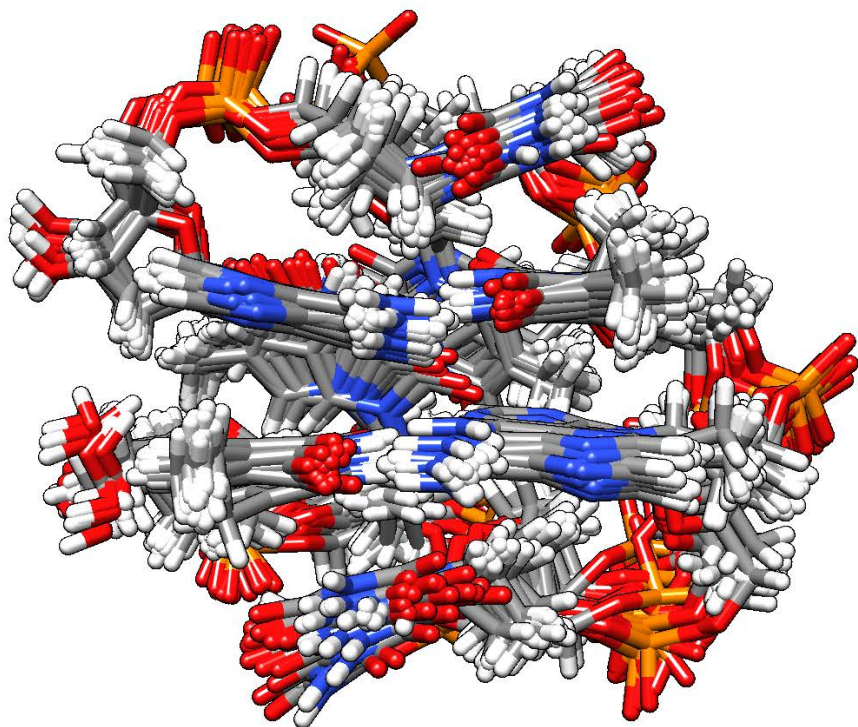
-Mg<sup>2+</sup> deviates from NMR structure: re-refine...



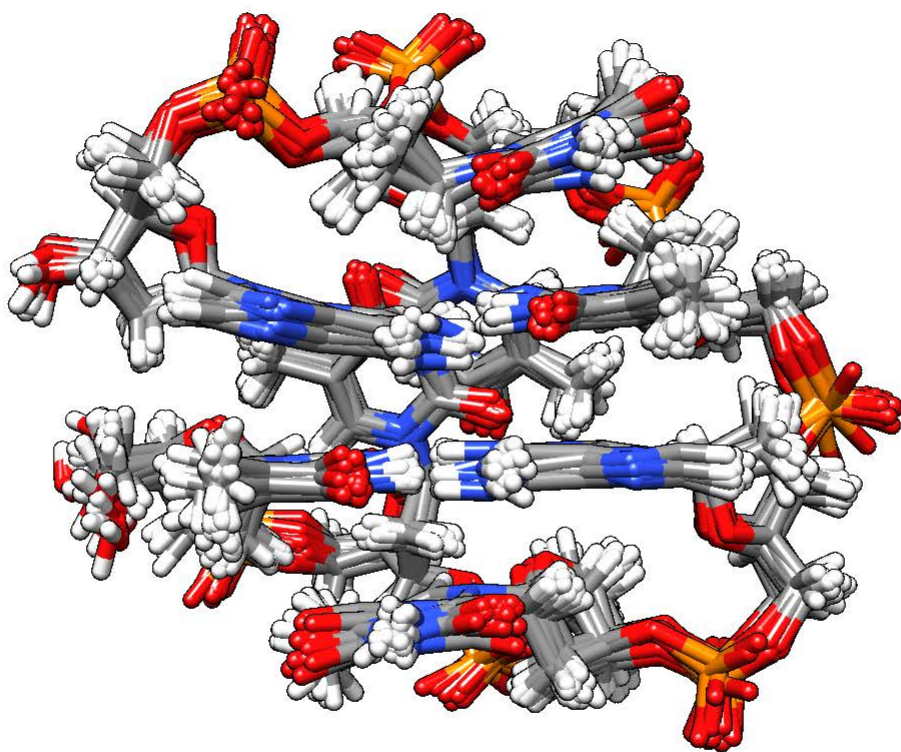
# NMR re-refinement TTTATTTA

- Starting from each of the 20 conformations → re-refine with bsc1/OL15 and opc/opc3 – with original restraint file (264 bond and angle restraints)
- Run form 100 ns, extract representative conformation from most populated cluster.

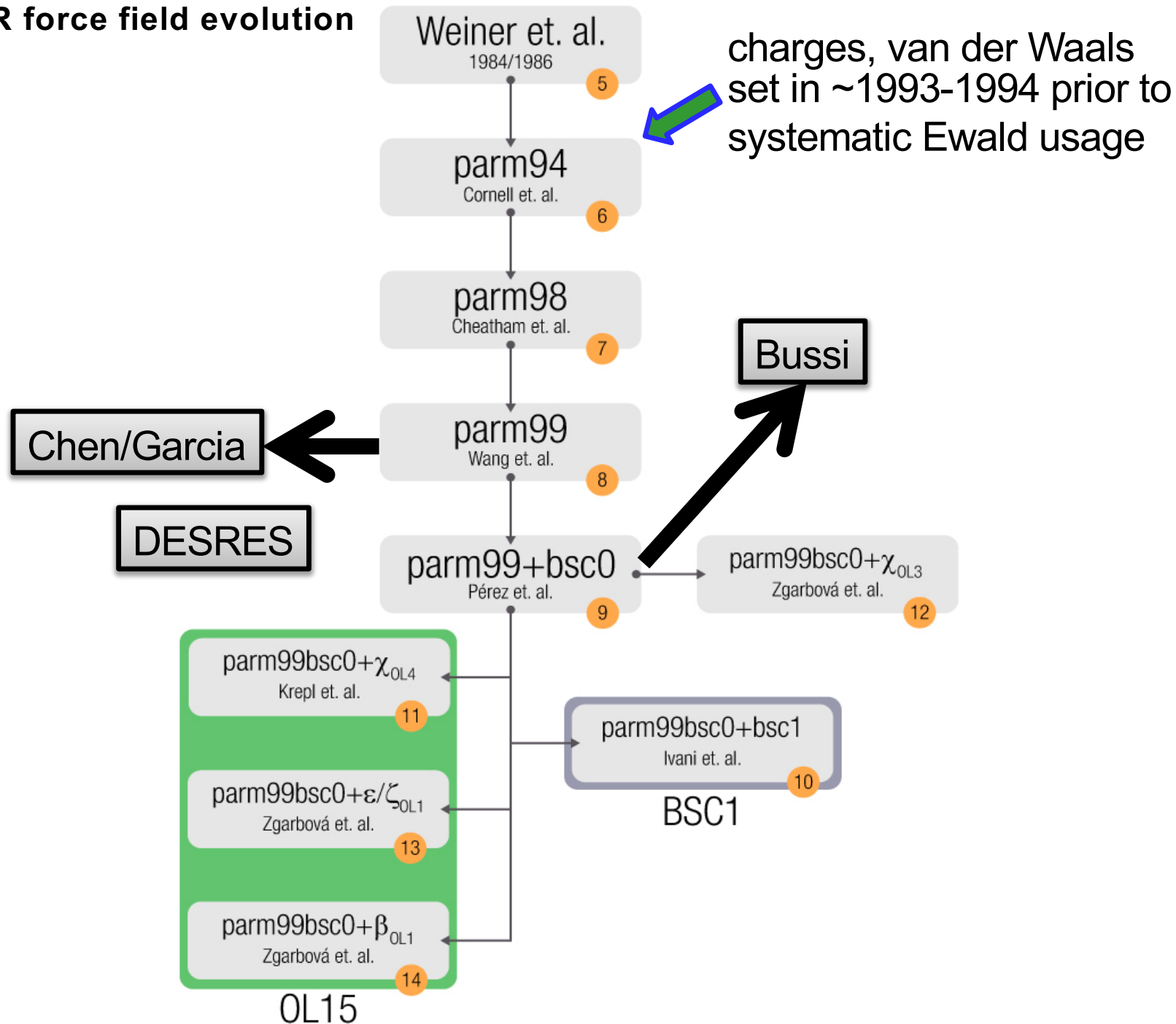
Pei Guo and Sik Lok Lam, JACS (2016)



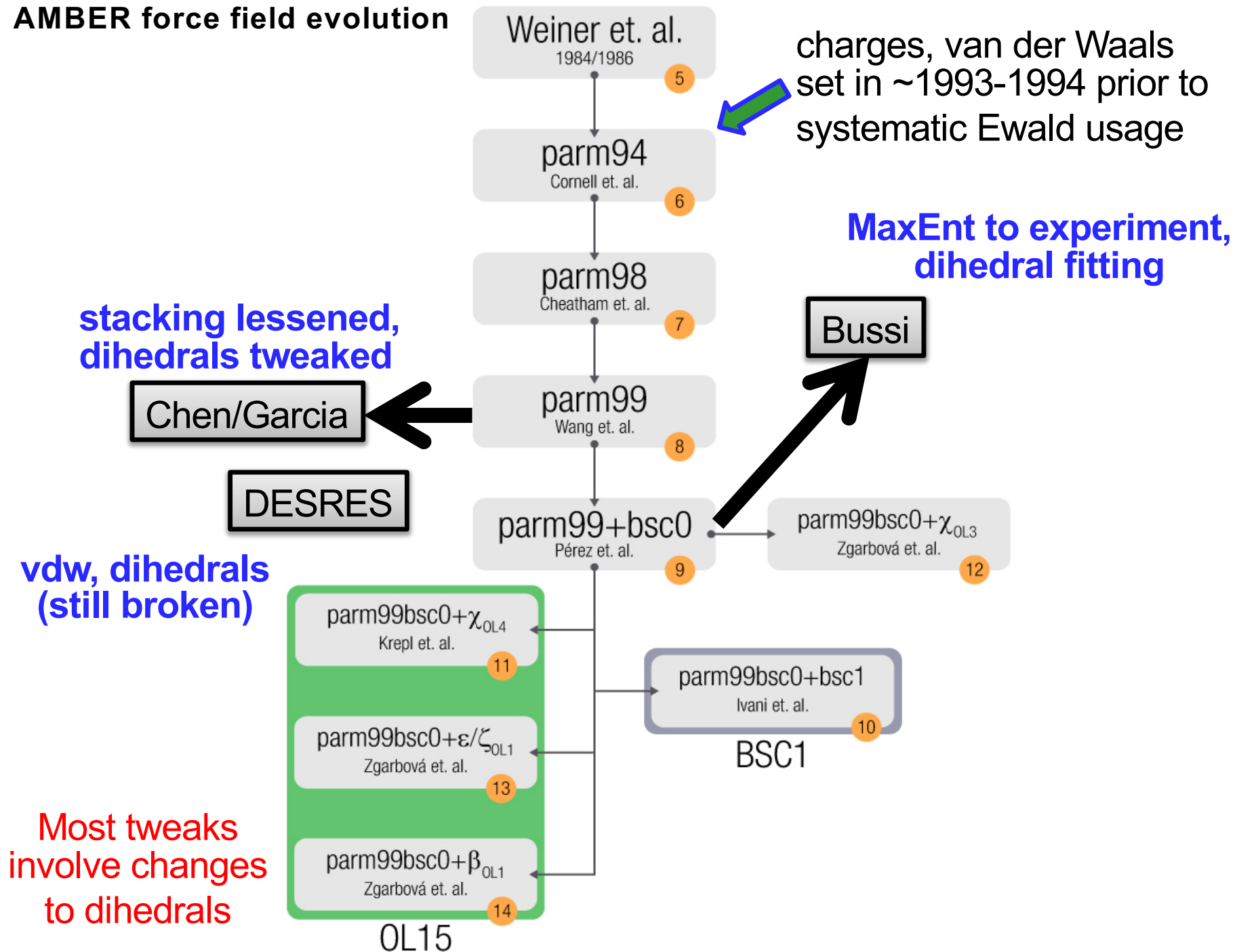
NMR original



# AMBER force field evolution



# AMBER force field evolution



# AMBER force field evolution

...we are finally starting to test Drude / polarizable (no results yet)

stacking lessened, dihedrals tweaked

Chen/Garcia

DESRES

vdw, dihedrals (still broken)

Most tweaks involve changes to dihedrals

Weiner et. al.  
1984/1986  
5

parm94  
Cornell et. al.  
6

parm98  
Cheatham et. al.  
7

parm99  
Wang et. al.  
8

parm99+bsc0  
Pérez et. al.  
9

parm99bsc0+ $\chi_{OL4}$   
Krepl et. al.  
11

parm99bsc0+ $\epsilon/\zeta_{OL1}$   
Zgarbová et. al.  
13

parm99bsc0+ $\beta_{OL1}$   
Zgarbová et. al.  
14

OL15

5

6

7

8

9

11

13

14

charges, van der Waals set in ~1993-1994 prior to systematic Ewald usage

MaxEnt to experiment, dihedral fitting

Bussi

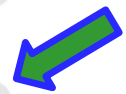
parm99bsc0+ $\chi_{OL3}$   
Zgarbová et. al.  
12

parm99bsc0+bsc1  
Ivani et. al.  
10

BSC1

OPC water model, phosphate modifications, sugar O's, O2' mods

...



**What is needed to properly set-up, run, assess and validate simulations of nucleic acids aimed at elucidating the “converged” conformational ensemble?**

**Initial conditions:**

- **starting structures, set-up (force fields, ions, water), equilibration?**

**“Production” molecular dynamics**

- **multiple independent runs and/or application of multiple types of enhanced sampling methods**

**ensembles,  
T-REMD,  
H-REMD,  
multidimensional REMD (T/H)**

**We can—using very long molecular dynamics (MD) simulations or even better using multidimensional replica exchange MD (M-REMD)—converge the conformational ensembles of various nucleic acids:**

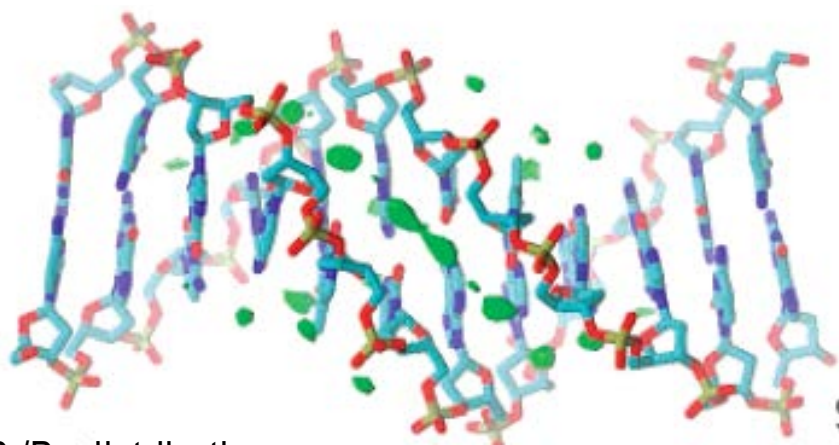
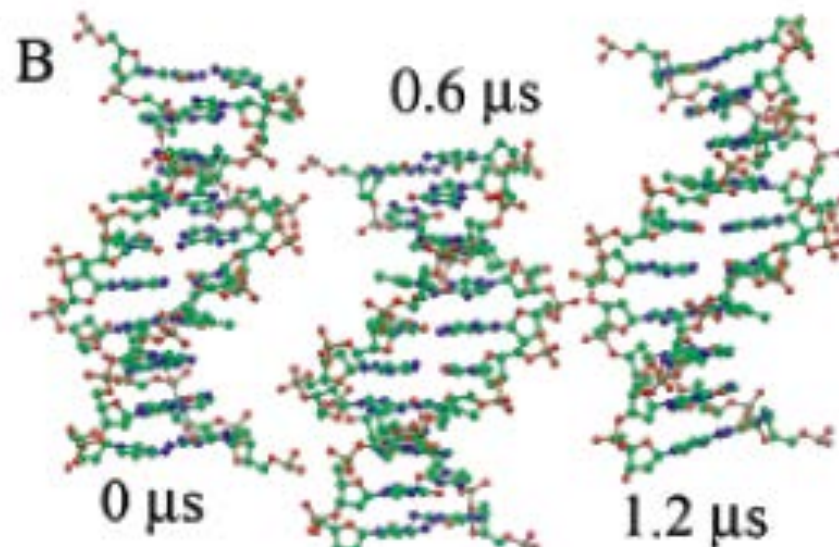
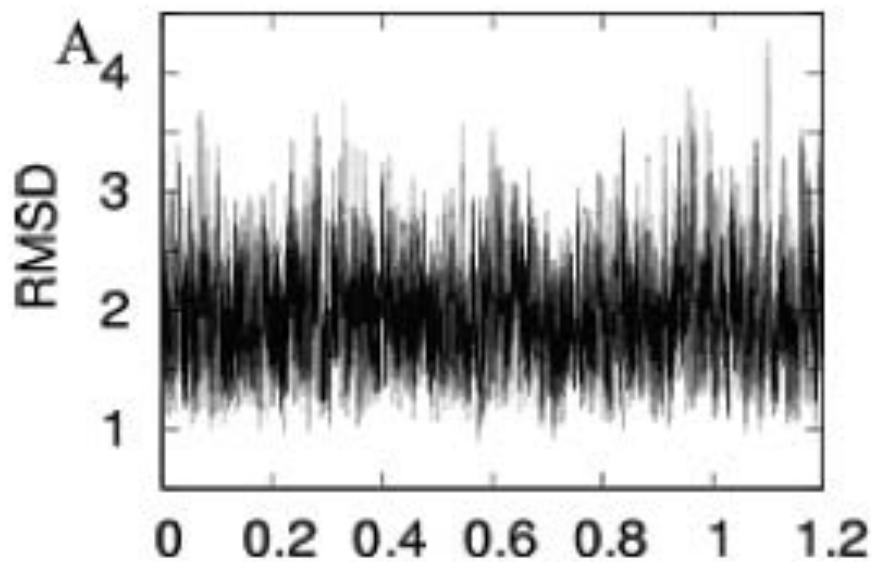
- duplexes
- dinucleotides
- tetranucleotides
- tetraloops (UUCG, GNRA, ...)
- mini-dumbbells (CCTGCCTG, TTTATTTA)
- *Soon:* NMR structures that are “dynamic”, e.g. UUCG, TAR, HIV SL1, A-loop, AAAA tetraloop, ...



# Dynamics of B-DNA on the Microsecond Time Scale

J. AM. CHEM. SOC. 2007,  
129, 14739–14745

Alberto Pérez.<sup>†,‡</sup> F. Javier Luque.<sup>§</sup> and Modesto Orozco\*.<sup>†,‡,||</sup>



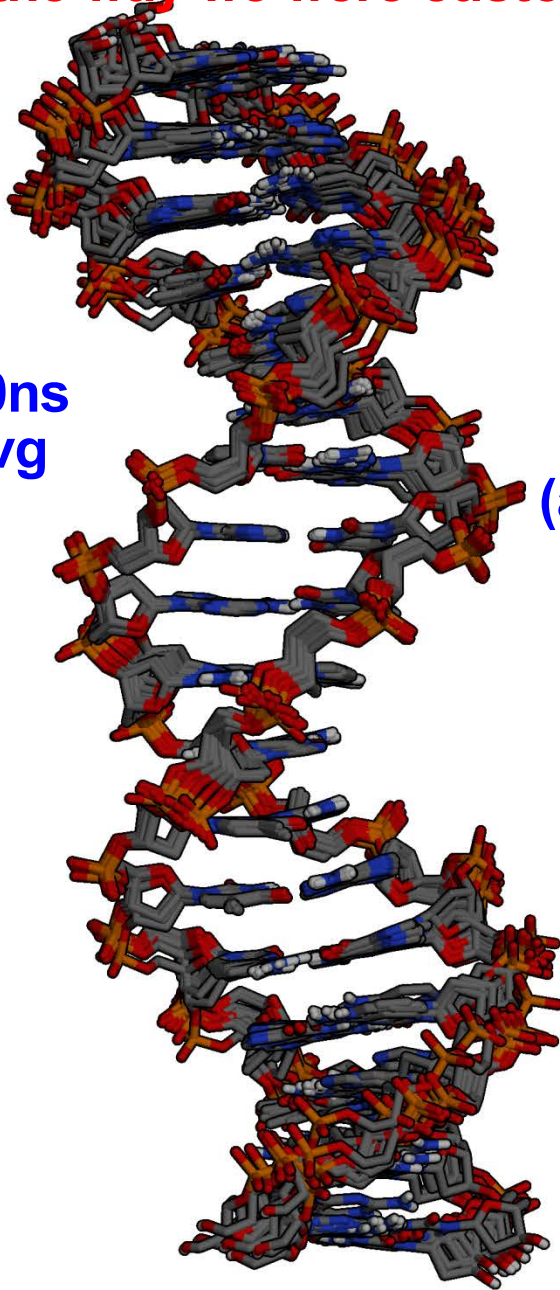
B<sub>I</sub>/B<sub>II</sub> distributions  
still changing



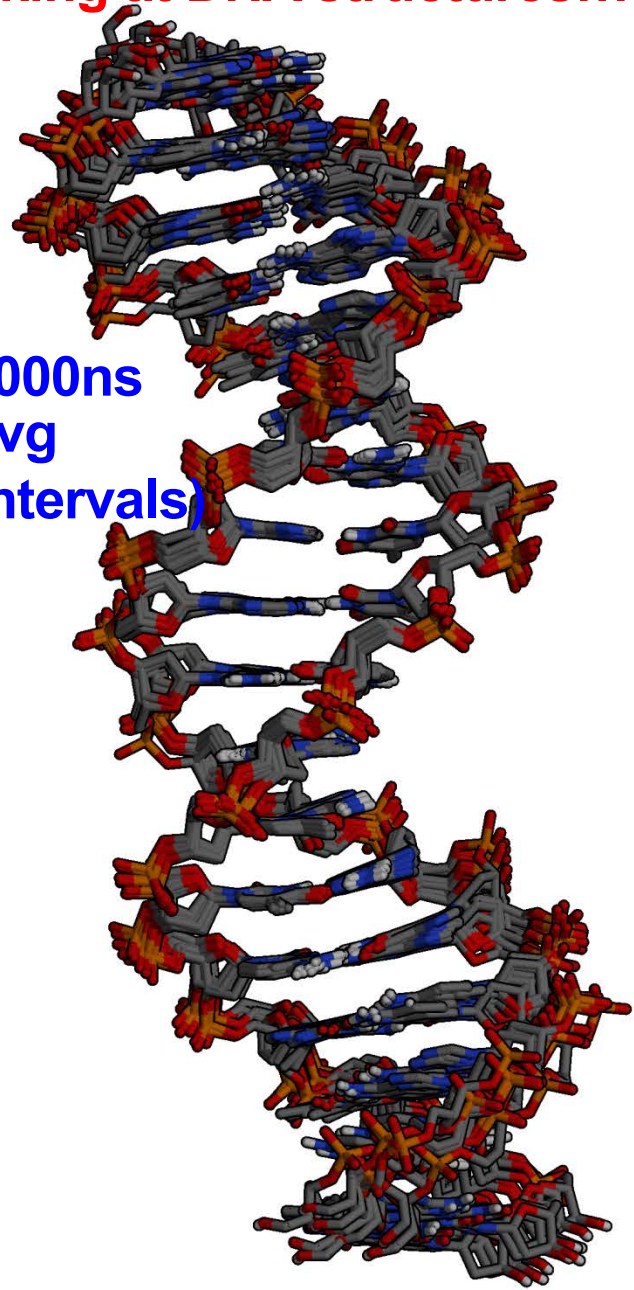
**Convergence? Not yet...**

...the way we were customarily looking at DNA structures...

abc, 50ns  
5ns avg



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

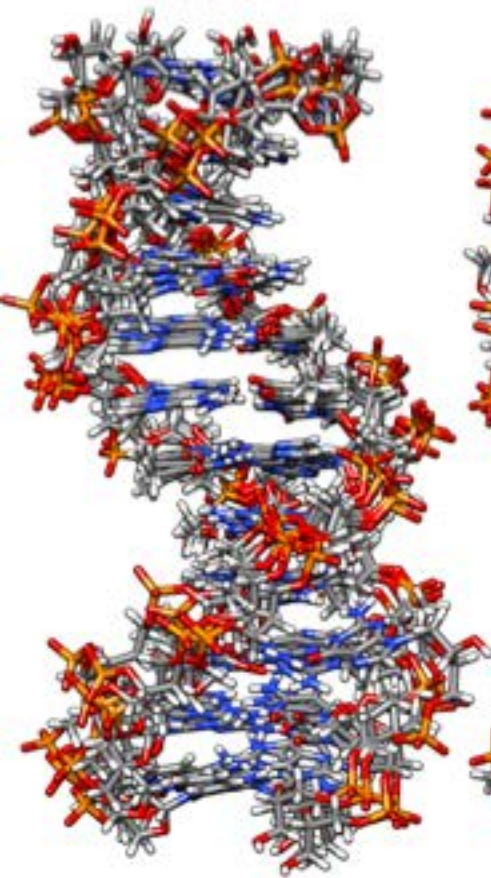


Where most “simulators” stop...



**1  $\mu$ s average structures**

net



200mM



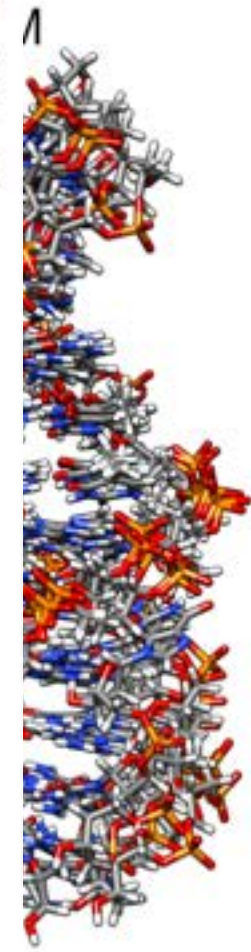
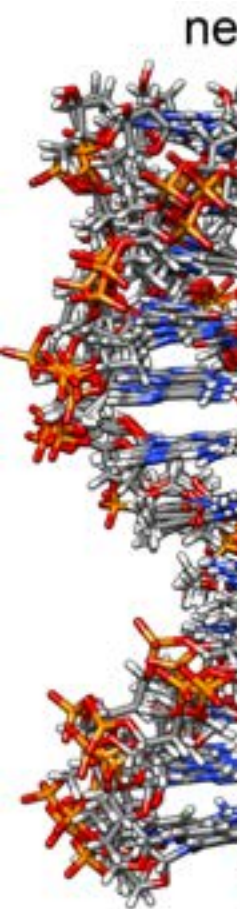
1M



5M



**Table 1.** RMS deviation values (Å) for the net neutralizing simulations and alternate conditions. A 1  $\mu$ s average structure was calculated for each case and used as reference. All frames in the trajectory were considered for the RMS calculation. Inner residue values correspond to the RMSD using residues 3 to 10 and 15 to 22 (i.e. omitting the two terminal base pairs on each end of the helix). The Hawkings, Cramer and Truhlar pairwise generalized Born model was used for the GB calculations[43].



	All residues		Inner residues	
		Std. Dev.		Std. Dev.
No salt	2.28	0.59	1.40	0.20
Li	2.08	0.51	1.62	0.31
Na	1.94	0.36	1.39	0.23
K	1.98	0.58	1.43	0.28
Rb	1.93	0.38	1.41	0.26
Cs	2.02	0.38	1.41	0.27
ff94bsc0	3.18	0.52	1.57	0.38
ff98bsc0	2.09	0.35	1.84	0.22
ff99bsc0	2.18	0.46	1.46	0.31
GB	4.19	0.71	3.46	0.75

# Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas

Bernhard Knapp,<sup>\*,†,‡,§</sup> Luis Ospina,<sup>‡,§</sup> and Charlotte M. Deane<sup>‡</sup>

<sup>†</sup>Bioinformatics and Immunoinformatics Research Group, Department of Basic Sciences, International University of Catalonia, 08195 Barcelona, Spain

<sup>‡</sup>Protein Informatics Group, Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom

<sup>§</sup>Alliance Manchester Business School, University of Manchester, Manchester M13 9SS, United Kingdom

## Supporting Information

**ABSTRACT:** Molecular simulations are a computational technique used to investigate the dynamics of proteins and other molecules. The free energy landscape of these simulations is often rugged, and minor differences in the initial velocities, floating-point precision, or underlying hardware can cause identical simulations (replicas) to take different paths in the landscape. In this study we investigated the magnitude of these effects based on 310 000 ns of simulation time. We performed 100 identically parametrized replicas of 3000 ns each for a small 10 amino acid system as well as 100 identically parametrized replicas of 100 ns each for an 827 residue T-cell receptor/MHC system. Comparing randomly chosen subgroups within these replica sets, we estimated the repro-

results from  
a single simulation



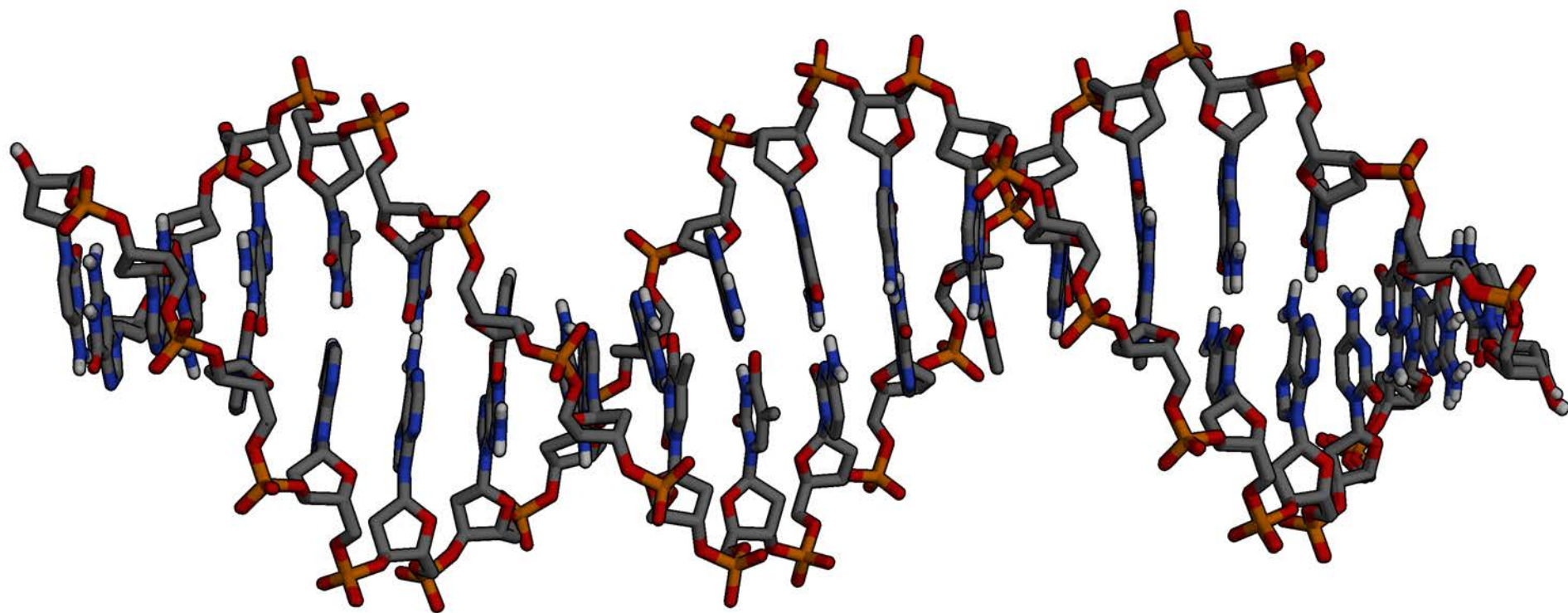
results from multiple replicas  
of the same simulation



## 5 “average” structures overlaid @

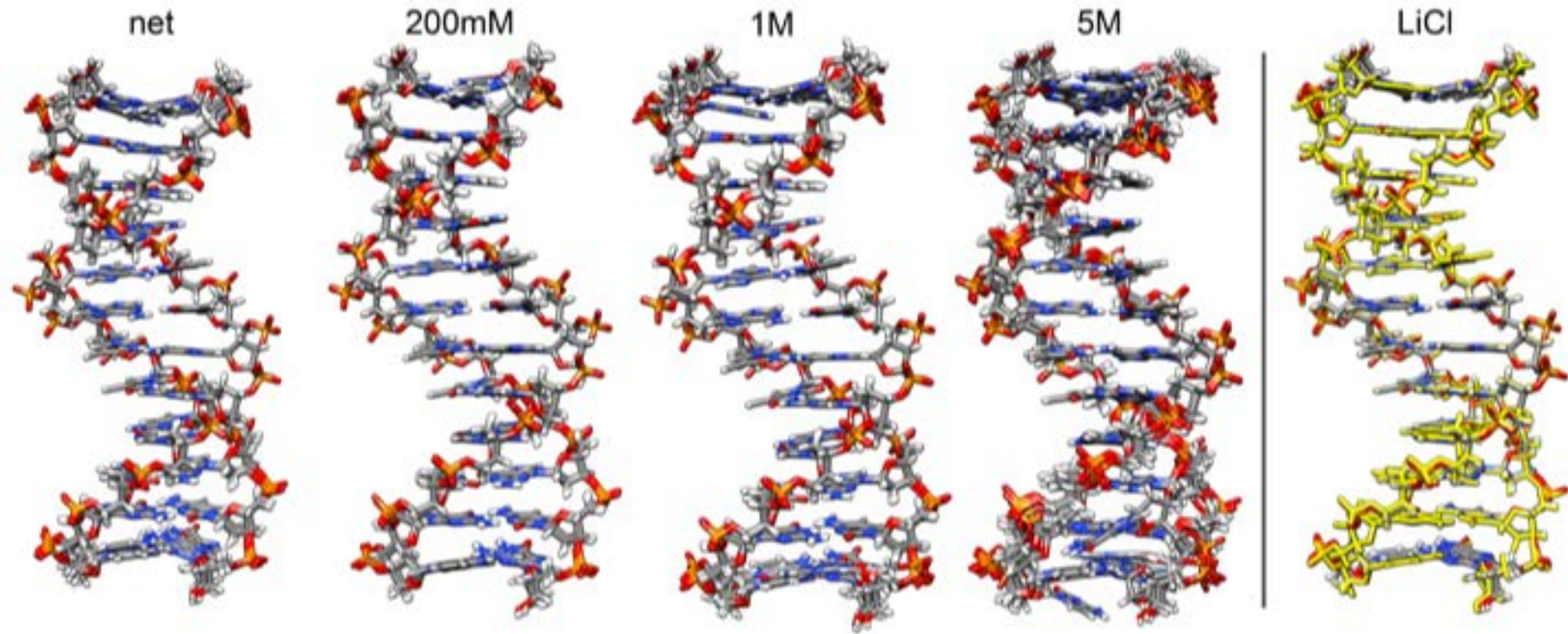
1.0-4.0  $\mu\text{s}$ , 1.5-4.5  $\mu\text{s}$ , 2.0-5.0  $\mu\text{s}$ , 2.5-5.5  $\mu\text{s}$ , 3.0-6.0  $\mu\text{s}$  ...

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



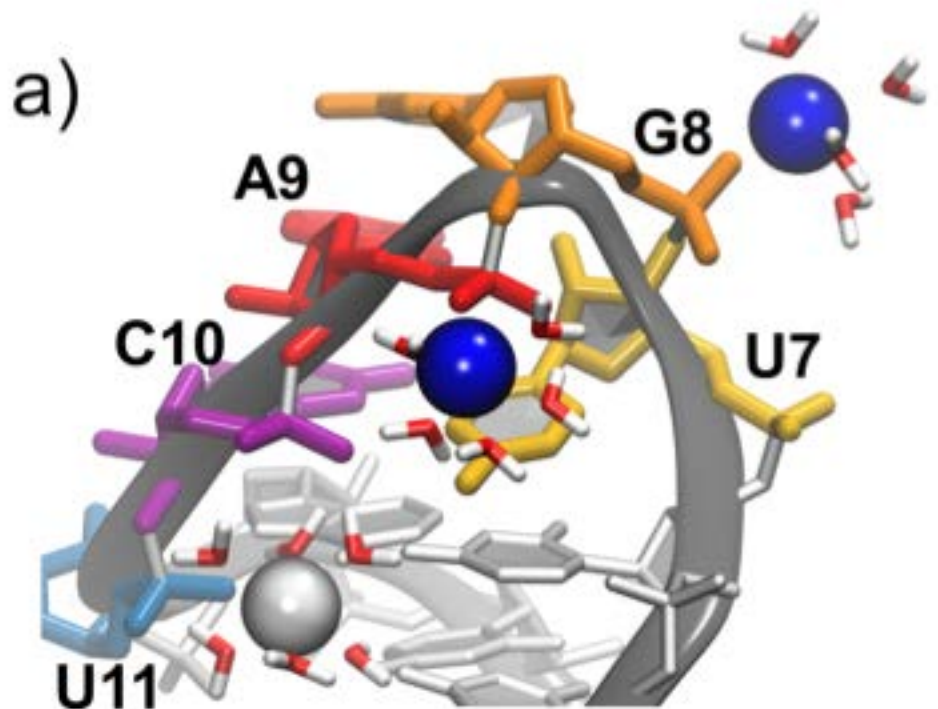
...then along came Anton and GPUs (BW)

# 10 $\mu$ s average structures



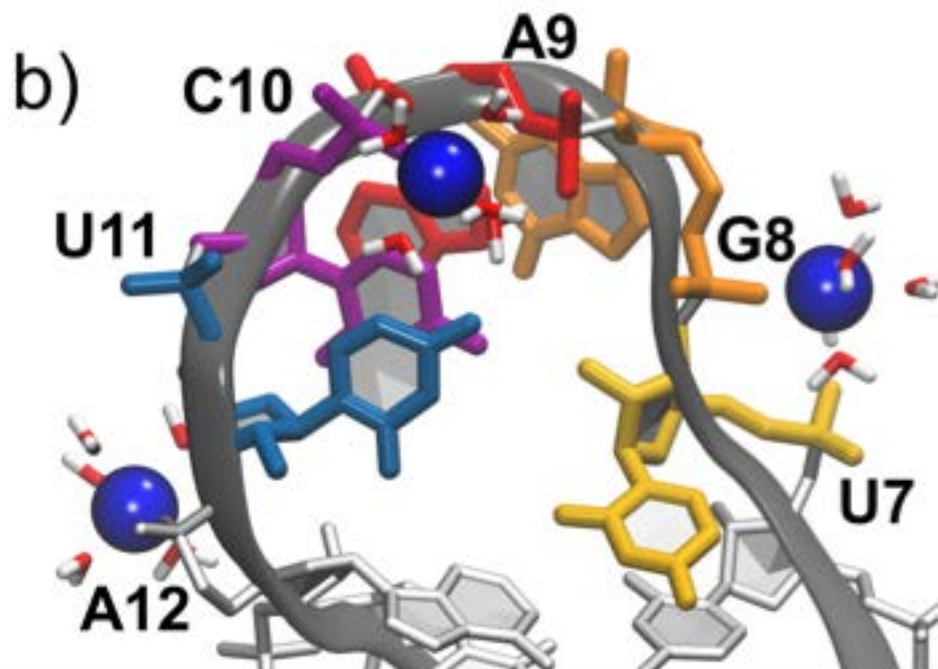
Little influence of salt concentration or identity, except groove narrowing at high salt (with current AMBER force fields)

OK 😊



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

trapped  
for ms



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



# What is needed to properly set-up, run, assess and validate simulations of nucleic acids aimed at elucidating the “converged” conformational ensemble?

## Initial conditions:

- starting structures, set-up (force fields, ions, water), equilibration?

## “Production” molecular dynamics

- multiple independent runs and/or application of multiple types of enhanced sampling methods

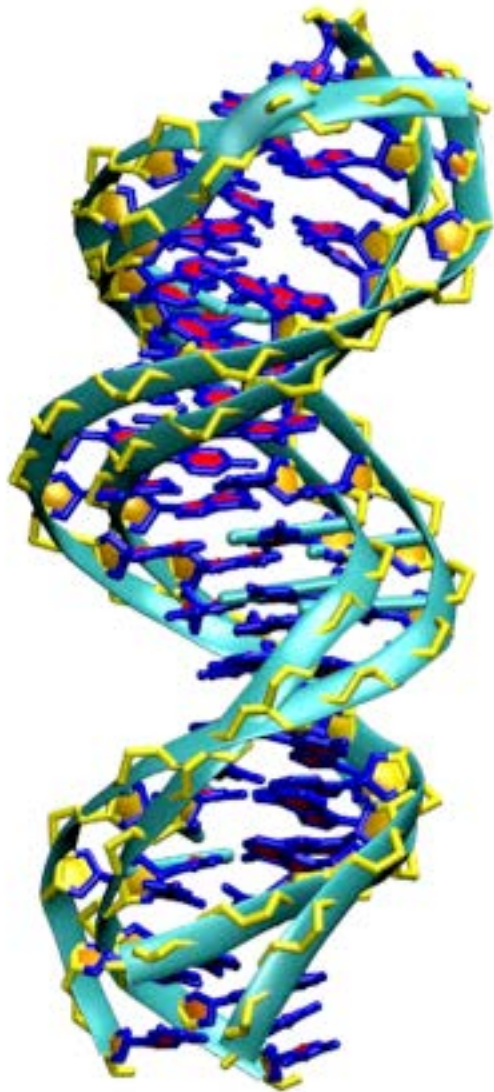
## When are you “done”?

- assessing convergence – measures of structure & dynamics

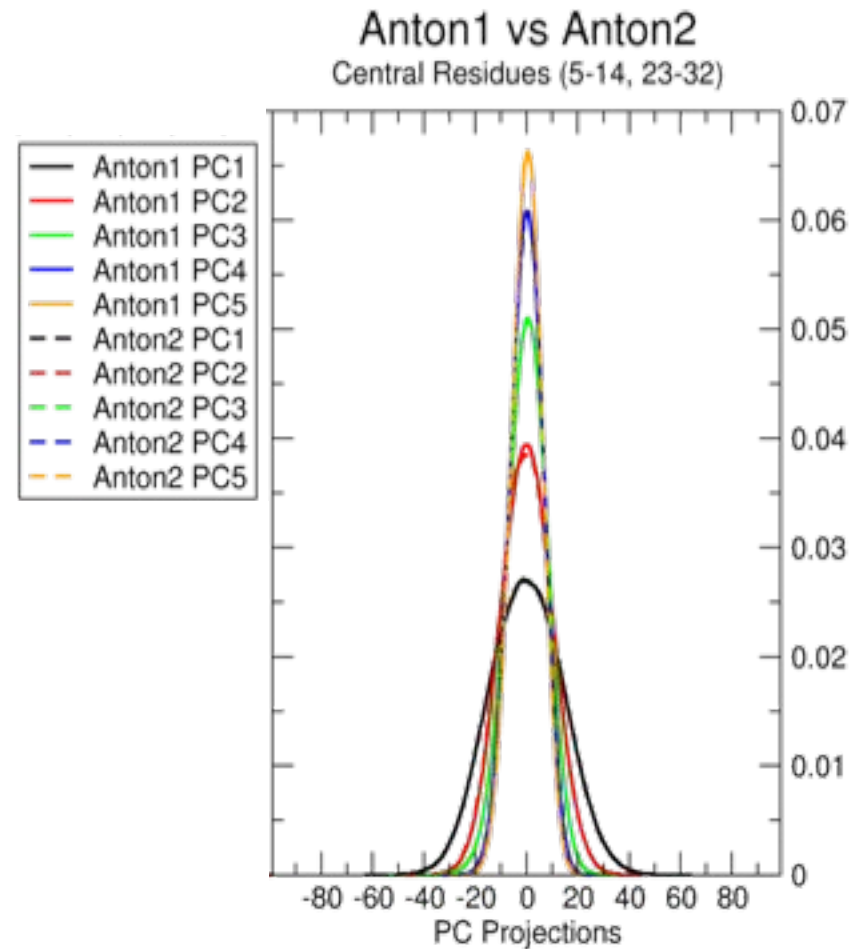
**“combined” clustering**

**“combined” PCA**

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



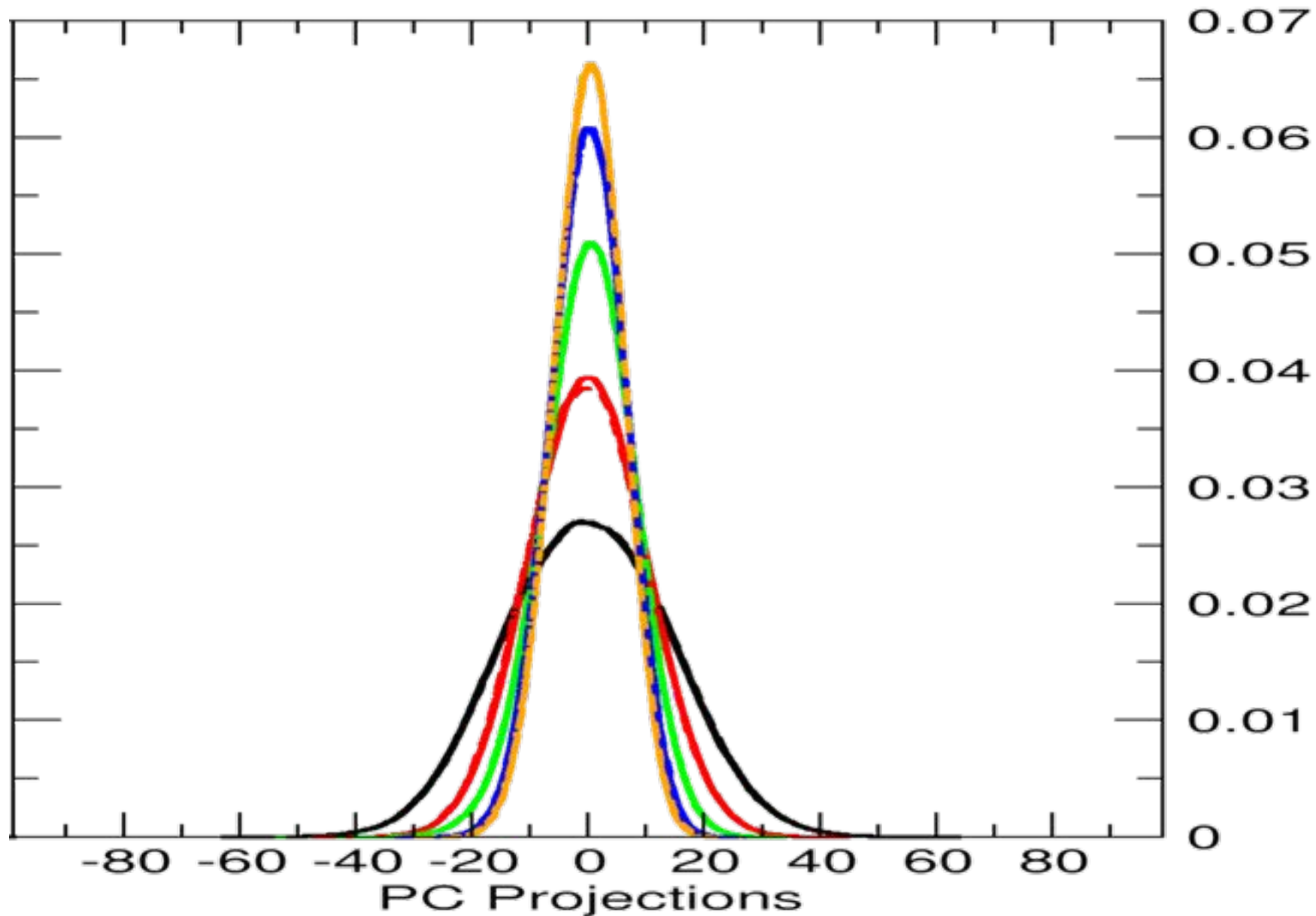
*Visualization of the first two (dominant) modes of motion*



*Overlap of modes from independent simulations (internal helix)*

# Anton1 vs Anton2

Central Residues (5-14, 23-32)



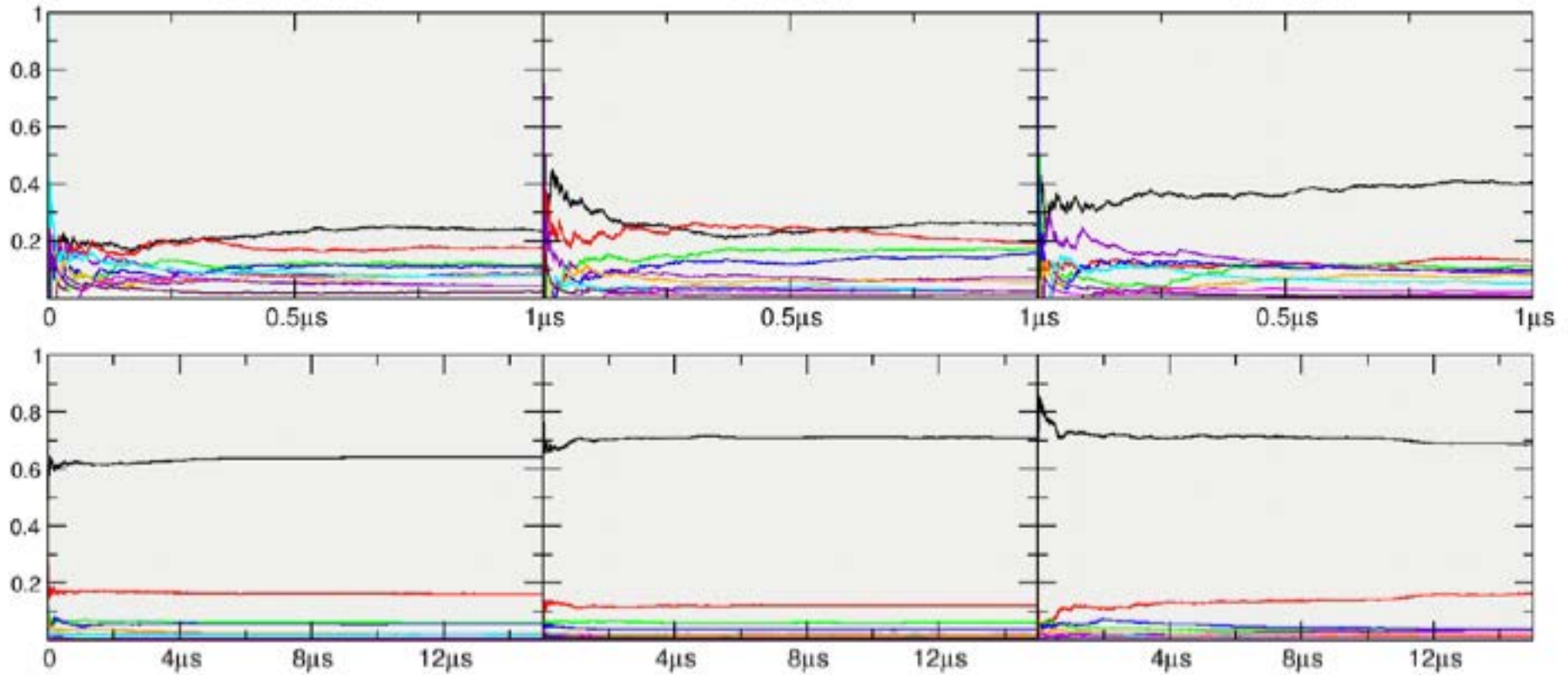


# cluster populations vs. time

200mM - NaCl

1M - NaCl

5M - NaCl



## What we have now in CPPTRAJ...

- MPI || across files
- MPI || across ensembles (independent sets of simulations)
- OpenMP for time consuming tasks (pairwise distance calculations)
- GPU Cuda for “most” time consuming tasks
  
- Python interface (pytraj)

### Newer stuff:

- calcstates (way to define “states“ from data) and do lifetimes, transition rates, ...
- Lennard Jones PME (library from Andy Simonett, NIH)
- data set caching to disk
- atom-mapping, best fit (lower) RMSD with symmetric-RMSD

Software News and Updates

## Parallelization of CPPTRAJ enables large scale analysis of molecular dynamics trajectory data

Daniel R. Roe , Thomas E. Cheatham III

First published: 03 October 2018 | <https://doi.org/10.1002/jcc.25382>

Contract Grant sponsor: Office of Advanced Cyberinfrastructure; Contract Grant number: 1443054

Contract Grant sponsor: Division of Chemistry; Contract Grant number: 1266307

Contract Grant sponsor: NSF; Contract Grant number: ACI-1443054 Run on NCSA Blue Waters using 8, 60, or 192 nodes, 1 process per node, with the trajectory data stored on a Lustre parallel file system.



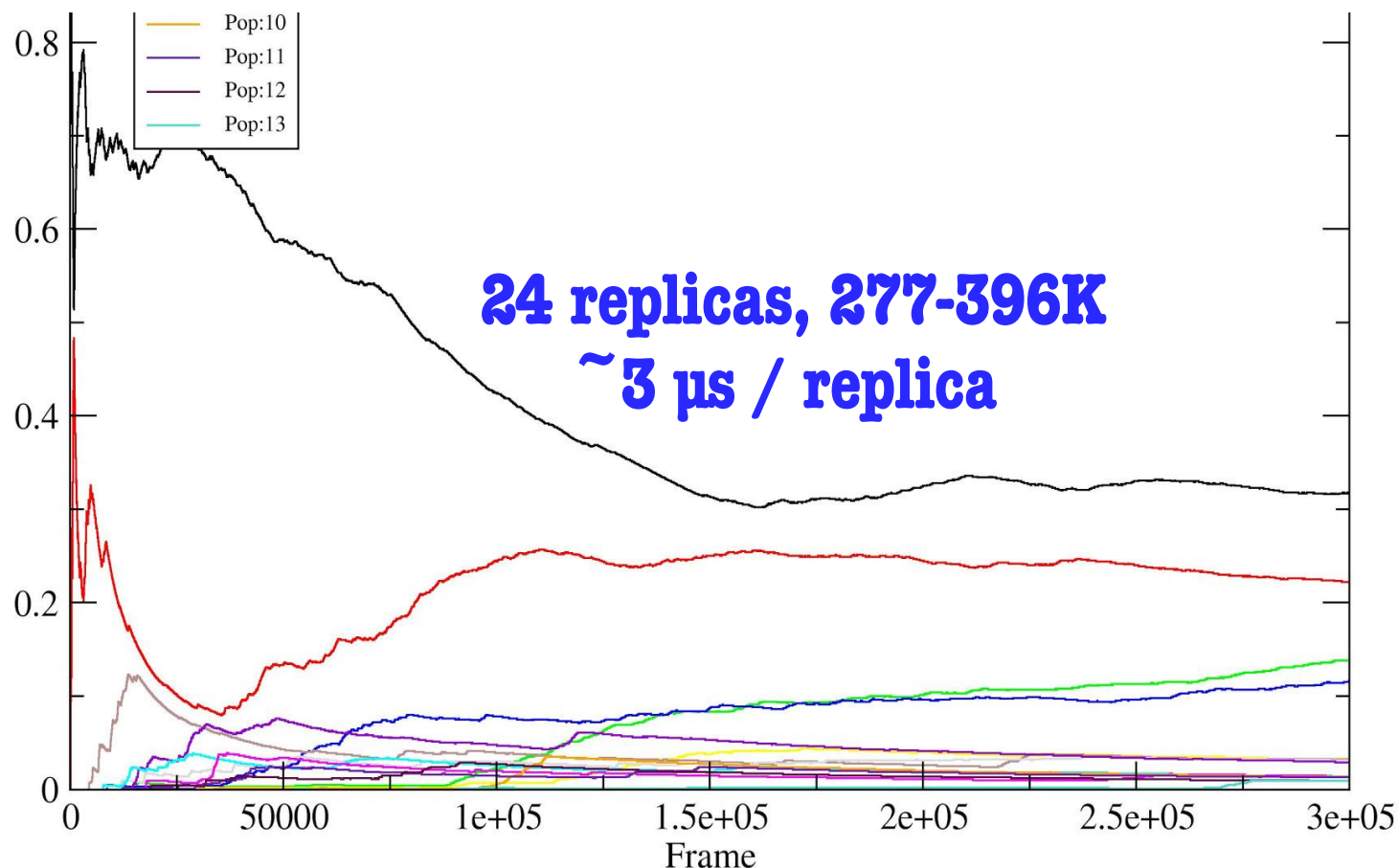
**Volume 39, Issue 25**

September 30, 2018

Pages 2110-2117

## Other issues:

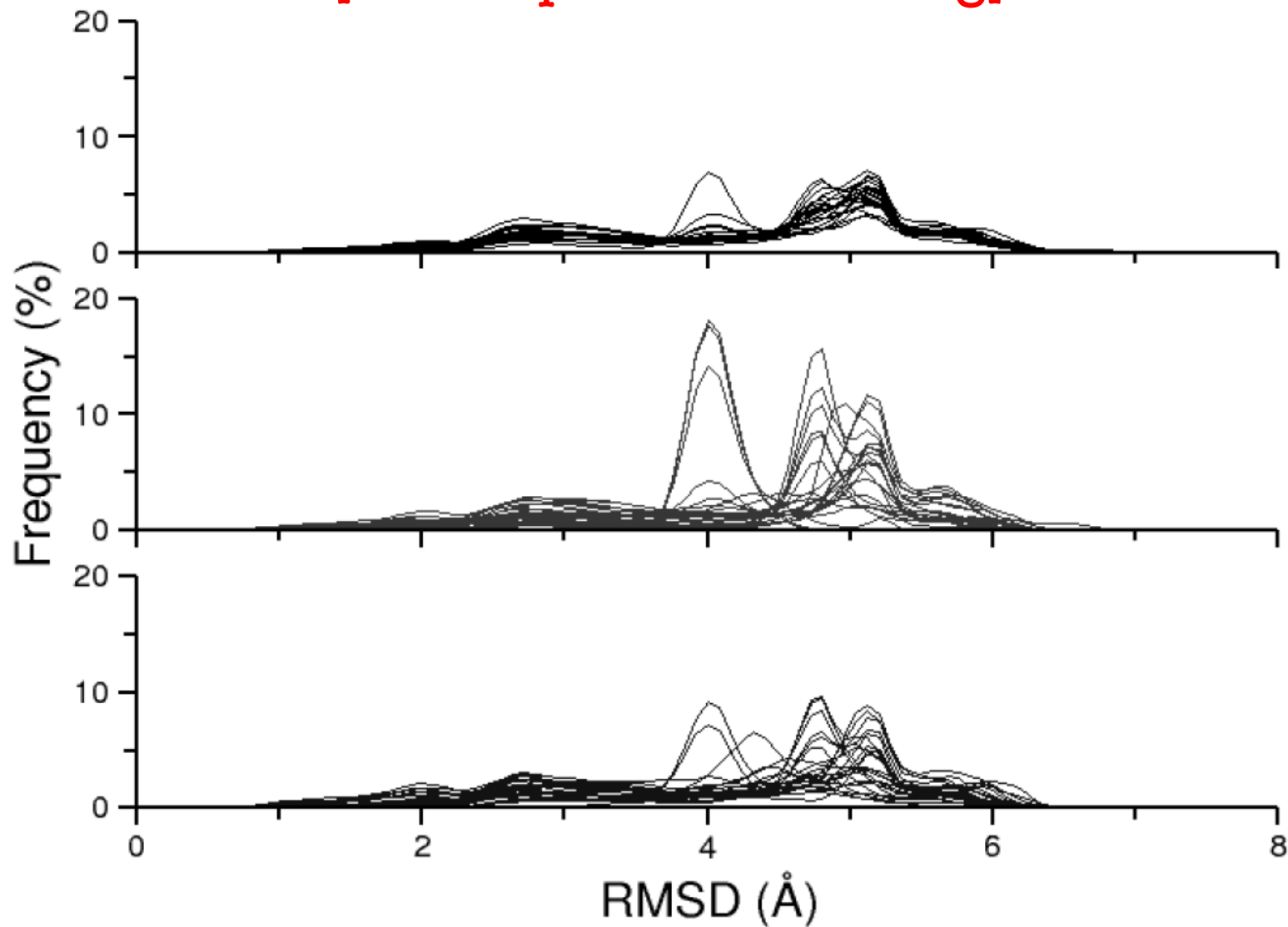
- **T-REMD still not “fully” converged (depending on def.)**



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



RMSd profiles per replica (they should be the same)  
[no temperature sorting]



# What is needed to properly set-up, run, assess and validate simulations of nucleic acids aimed at elucidating the “converged” conformational ensemble?

## Initial conditions:

- starting structures, set-up (force fields, ions, water), equilibration?

## “Production” molecular dynamics

- multiple independent runs and/or application of multiple types of enhanced sampling methods

## When are you “done”?

- assessing convergence – measures of structure & dynamics

## How to validate?

- This is tricky: What should the populations of minor conformations be?

**We can—using very long molecular dynamics (MD) simulations or even better using multidimensional replica exchange MD (M-REMD)—converge the conformational ensembles of various nucleic acids:**

- duplexes
- dinucleotides
- tetranucleotides
- tetraloops (UUCG, GNRA, ...)
- mini-dumbbells (CCTGCCTG, TTTATTTA)
- *Soon:* NMR structures that are “dynamic”, e.g. UUCG, TAR, HIV SL1, A-loop, AAAA tetraloop, ...

**We can assess various force fields, re-weight to experimental observables, and parameter scan various changes to the underlying potentials to ultimately capture the influence on the conformational ensemble...**

# We can ...re-weight to experimental observables

SCIENCE ADVANCES | RESEARCH ARTICLE

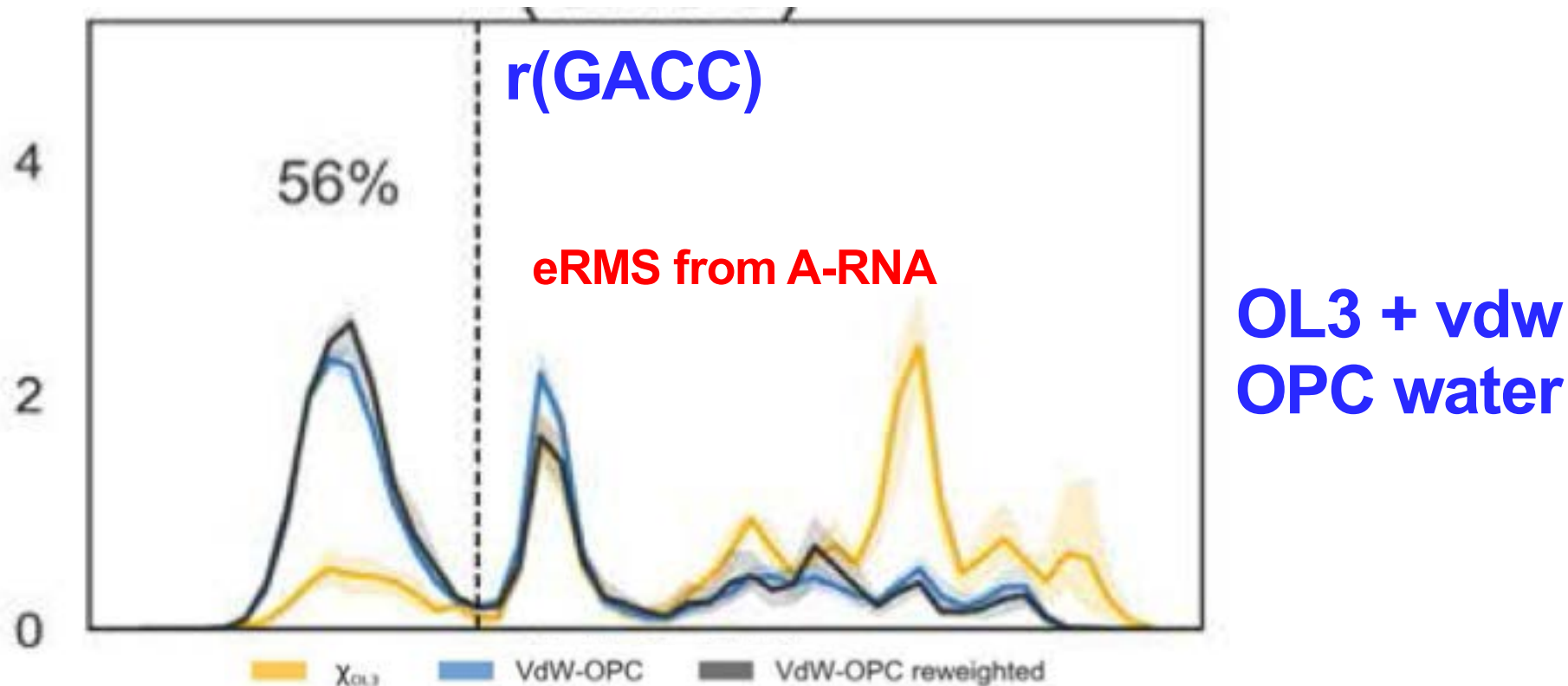
BIOCHEMISTRY

*Sci. Adv.* 2018;4:eaar8521

18 May 2018

## Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations

Sandro Bottaro,<sup>1\*</sup> Giovanni Bussi,<sup>2</sup> Scott D. Kennedy,<sup>3</sup> Douglas H. Turner,<sup>4</sup> Kresten Lindorff-Larsen<sup>1\*</sup>



QM on crystals of bases, RESP on dinucleotides, small organics,  
parameter scanning, open-FF consortium, M-BAR re-weighting

**“new”  $q$ ,  $\epsilon$ ,  $r^*$**

different validations?

**alternative sequence  
tetranucleotides**

MD ensembles populating  
weird structures  
subject to NMR

asynchronous, adaptable

**M-REMD**

Temperature, Hamiltonians:  
various force fields,  
reduce dihedral force constants,  
aMD, parameter scanning

Experimentally verifiable

**models**

dinucleotides;  
GACC, AAAA, UUUU,  
CCCC, CAAU ;  
UUCG, GNRA, CUUG  
tetraloops ;  
TTTATTTA dumbbell

**Force field**

**improvement?**

Move to dynamic, multiple  
minimum RNA structures  
with strong NMR: TAR,  
ribosomal A-site, HIV  
SL1, ...

compare to experiment

**NMR, MaxEnt**

J coupling, NOEs, uNOES,  
RDCs, relaxation, ...



steered

**CPPTRAJ**

analysis:  
replica round-trip times,  
exchange rate,  
convergence of cluster  
populations  
and principle modes,  
“seeding” new conformers,  
thermodynamic properties

If these work, move to:  
riboswitches, RNA  
thermometers, xrRNA

<https://amberhub.chpc.utah.edu/>

# AMBER-Hub

[Home](#) [CPPTRAJ](#) [CPPTRAJ cookbook](#) [Tutorials](#) [About](#)

## AMBER and CPPTRAJ examples and recipes

A place created by users of the AMBER suite of biomolecular simulation programs to share, enrich and contribute to the learning and use of this tool. A computational chemist cookbook for recipes on how to use the codes provided in the AMBER tools package. We provide an easy-to-use place to learn the tool CPPTRAJ, the default package to perform analysis of trajectory information generated by the AMBER programs and suit of force fields.



Start here

CPPTRAJ introductory information to perform analysis



CPPTRAJ Manual

Overview of CPPTRAJ capabilities



CPPTRAJ cookbook

Collection of CPPTRAJ recipes and one-liners to perform analysis



AMBER examples

Collection of AMBER tutorials and procedures to perform molecular dynamics simulations

**Rodrigo Galindo (Research Assistant Professor, U Utah)**

## Recipes index.

Combining multiple trajectory files into a single trajectory and remove water molecules to save space

Reading the OUT files and plotting different properties

Generating histograms with CPPTRAJ

Histogram analysis of dihedral angle distributions

Hydrogen bond analysis between a protein and a small molecule

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

**\$\$\$:**



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 2013-2018

## Computer time:



D E Shaw Research

“Anton”  
(3 past awards)



XRAC MCA01S027 ~12M GPU hours per year  
~10M core hours

!!!



~3M hours

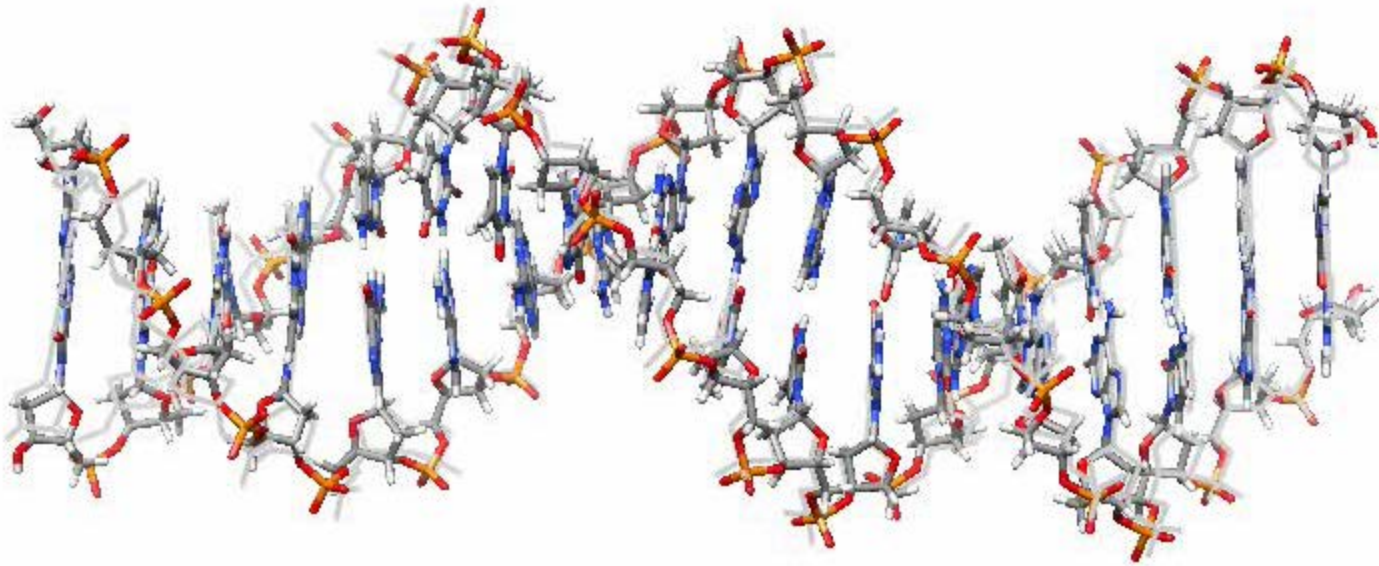


# Products

- 3 PRAC awards (2011-2018), 1 Ebola RAPID
- 50+ Cheatham group publications, 2013-6/2019
- GPU-accelerated Amber 14, Amber 16, Amber 18/19
- multi-dimensional replica exchange (M-REMD)
- 4 levels of parallelism in CPPTRAJ (molecular dynamics trajectory analyses – ensemble, file/analyses, OpenMP, CUDA) **[JCC paper published]**
- method validation (Anton vs. AMBER vs. GROMACS vs. CHARMM)
- re-refined NMR structures, Mg-dependent structure
- hydrogen mass repartitioning
- reproducibility & convergence
- force field assessment / validation / optimization

# questions?

---



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