Practical considerations in running simulations and example applications

Nov. 14, 2016

JC Gumbart

Assistant Professor of Physics

Georgia Institute of Technology

gumbart@physics.gatech.edu

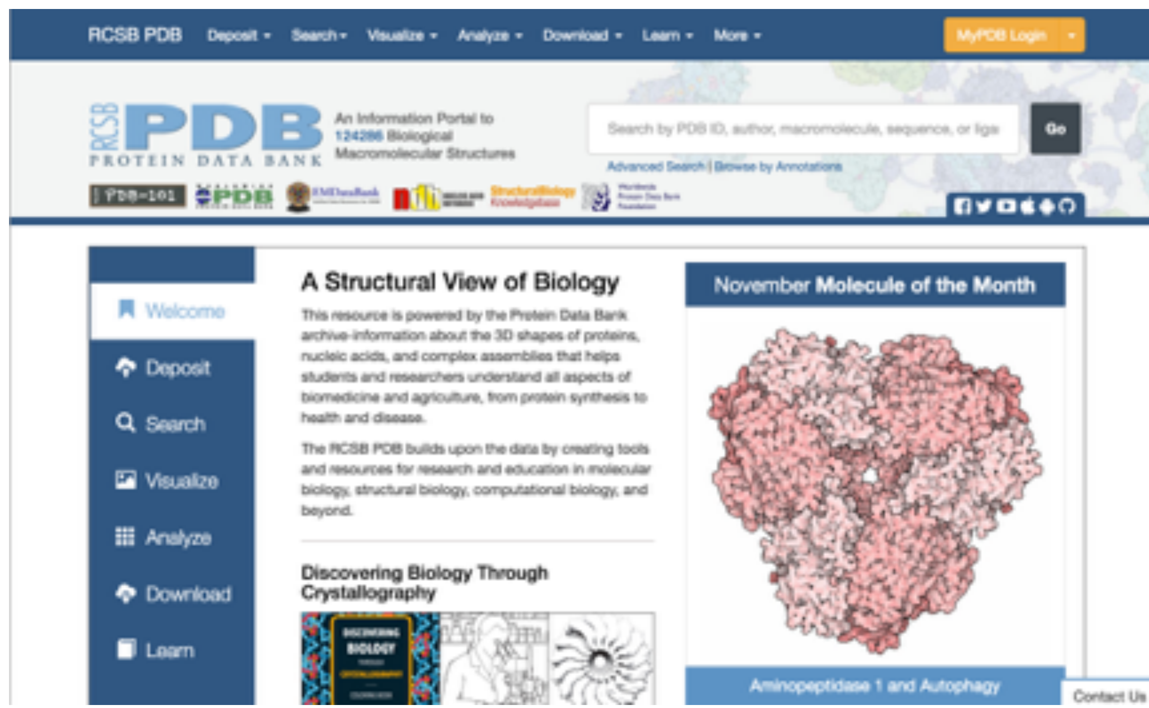
simbac.gatech.edu

PDB files *provide the structure and starting position*

- Simulations start with a crystal structure from the Protein Data Bank, in the standard PDB file format.
- PDB files contain standard records for species, tissue, authorship, citations, sequence, secondary structure, etc.
- We only care about the atom records…
  - – atom name (N, C, CA)
  - – residue name (ALA, HIS)
  - – residue id (integer)
  - – coordinates (x, y, z)
  - – occupancy (0.0 to 1.0)
  - – temp. factor (a.k.a. beta)
  - – segment id (6PTI)
- No hydrogen atoms!

(We must add them ourselves.)

http://www.rcsb.org/



Structure of a PDB file

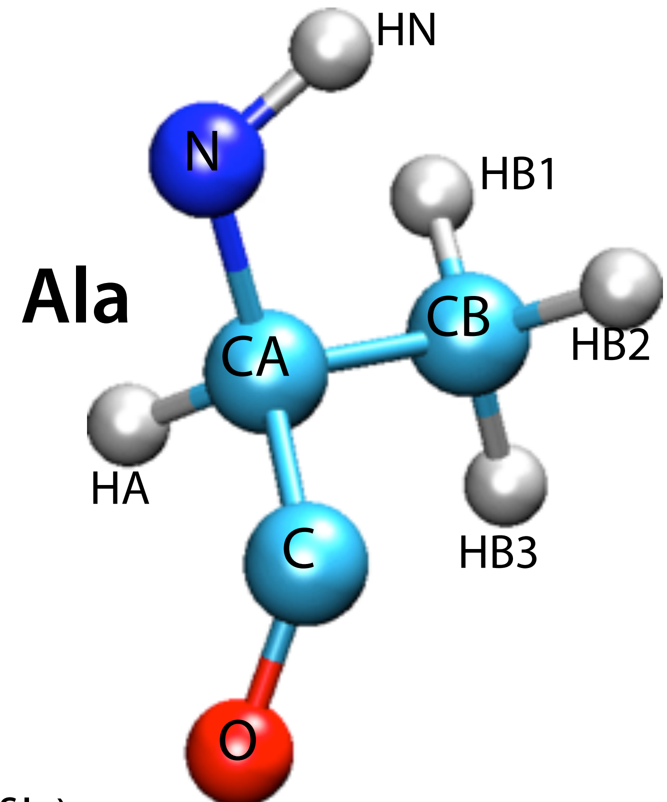
index · **name** · **resname** · **chain** · **resid** · **X** · **Y** · **Z** · **occupancy** · **beta** · **segname**

```
ATOM     22  N    ALA B   3      -4.073  -7.587  -2.708  1.00  0.00      BH
ATOM     23  HN   ALA B   3      -3.813  -6.675  -3.125  1.00  0.00      BH
ATOM     24  CA   ALA B   3      -4.615  -7.557  -1.309  1.00  0.00      BH
ATOM     25  HA   ALA B   3      -4.323  -8.453  -0.704  1.00  0.00      BH
ATOM     26  CB   ALA B   3      -4.137  -6.277  -0.676  1.00  0.00      BH
ATOM     27  HB1  ALA B   3      -3.128  -5.950  -0.907  1.00  0.00      BH
ATOM     28  HB2  ALA B   3      -4.724  -5.439  -1.015  1.00  0.00      BH
ATOM     29  HB3  ALA B   3      -4.360  -6.338   0.393  1.00  0.00      BH
ATOM     30  C    ALA B   3      -6.187  -7.538  -1.357  1.00  0.00      BH
ATOM     31  O    ALA B   3      -6.854  -6.553  -1.264  1.00  0.00      BH
ATOM     32  N    ALA B   4      -6.697  -8.715  -1.643  1.00  0.00      BH
ATOM     33  HN   ALA B   4      -6.023  -9.463  -1.751  1.00  0.00      BH
ATOM     34  CA   ALA B   4      -8.105  -9.096  -1.934  1.00  0.00      BH
ATOM     35  HA   ALA B   4      -8.287  -8.878  -3.003  1.00  0.00      BH
0000000000000000000000000000000000000000000000000000000000000000000000
        10        20        30        40        50        60        70
```

It is an ASCII, **fixed-width** file, which generally does not contain any connectivity information

PSF files *provide the topology and charges*

- Every atom in the simulation is listed.
- Provides all static atom-specific values:
  - atom name (N, C, CA)
  - atom type (NH1, C, CT1)
  - residue name (ALA, TRP)
  - residue id (integer)
  - segment id (6PTI)
  - atomic mass (in atomic mass units)
  - partial charge (in electronic charge units)
- What is not in the PSF file?
  - coordinates (dynamic data, initially read from PDB file)
  - velocities (dynamic data, initially from Boltzmann distribution)
  - force field parameters (non-specific, used for many molecules)



Structure of a PSF file

index **segname** **resid** **resname** **name** **type*** **charge*** **mass*** **unused**

```
 160 C    10   GLY  C    C      0.510000       12.0110          0
 161 C    10   GLY  O    O     -0.510000       15.9990          0
 162 C    11   ALA  N    NH1   -0.470000       14.0070          0
 163 C    11   ALA  HN   H      0.310000        1.0080          0
 164 C    11   ALA  CA   CT1    0.070000       12.0110          0
 165 C    11   ALA  HA   HB1    0.090000        1.0080          0
 166 C    11   ALA  CB   CT3   -0.270000       12.0110          0
 167 C    11   ALA  HB1  HA3    0.090000        1.0080          0
 168 C    11   ALA  HB2  HA3    0.090000        1.0080          0
 169 C    11   ALA  HB3  HA3    0.090000        1.0080          0
 170 C    11   ALA  C    C      0.510000       12.0110          0
 171 C    11   ALA  O    O     -0.510000       15.9990          0
 172 C    12   THR  N    NH1   -0.470000       14.0070          0
 173 C    12   THR  HN   H      0.310000        1.0080          0
 174 C    12   THR  CA   CT1    0.070000       12.0110          0
000000000000000000000000000000000000000000000000000000000000000000000000000
    10        20        30         40         50         60         70
```

Also an ASCII, **fixed-width** file, which does not contain coordinate information

Structure of a PSF file (extended)

index **segname** **resid** **resname** **name** **type*** **charge*** **mass*** **unused**

```
       82 A        212      THR      O        O        -0.510000      15.9990          0
       83 A        213      ALA      N        NH1      -0.470000      14.0070          0
       84 A        213      ALA      HN       H         0.310000       1.0080          0
       85 A        213      ALA      CA       CT1       0.070000      12.0110          0
       86 A        213      ALA      HA       HB1       0.090000       1.0080          0
       87 A        213      ALA      CB       CT3      -0.270000      12.0110          0
       88 A        213      ALA      HB1      HA3       0.090000       1.0080          0
       89 A        213      ALA      HB2      HA3       0.090000       1.0080          0
       90 A        213      ALA      HB3      HA3       0.090000       1.0080          0
       91 A        213      ALA      C        C         0.510000      12.0110          0
       92 A        213      ALA      O        O        -0.510000      15.9990          0
       93 A        214      GLU      N        NH1      -0.470000      14.0070          0
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
       10        20        30         40         50         60         70        80
```

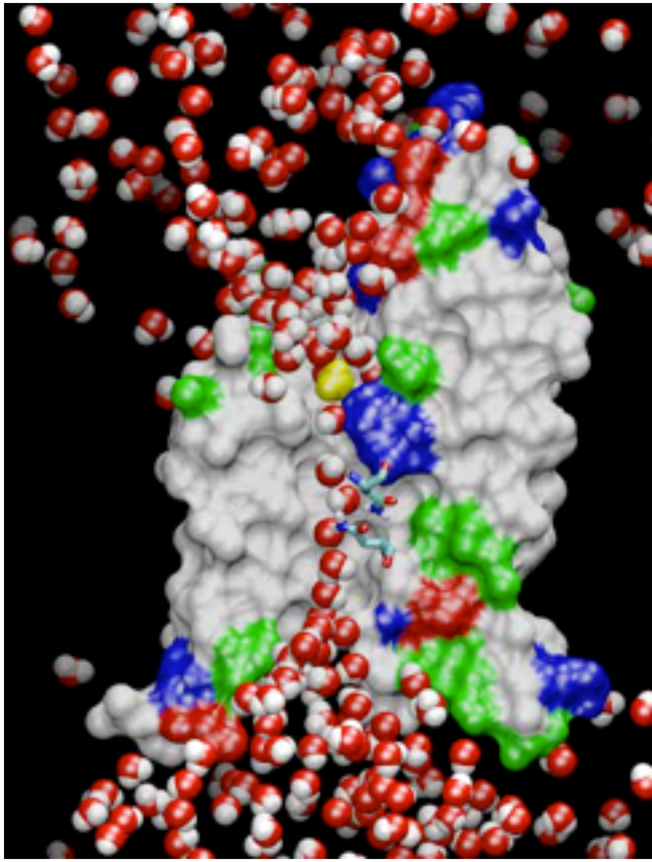
"Extended" format supports long atom types and names (> 4 characters)

Current versions of NAMD and VMD handle this automatically

`PSF EXT CMAP` "EXT" at the beginning of the file indicates extended format

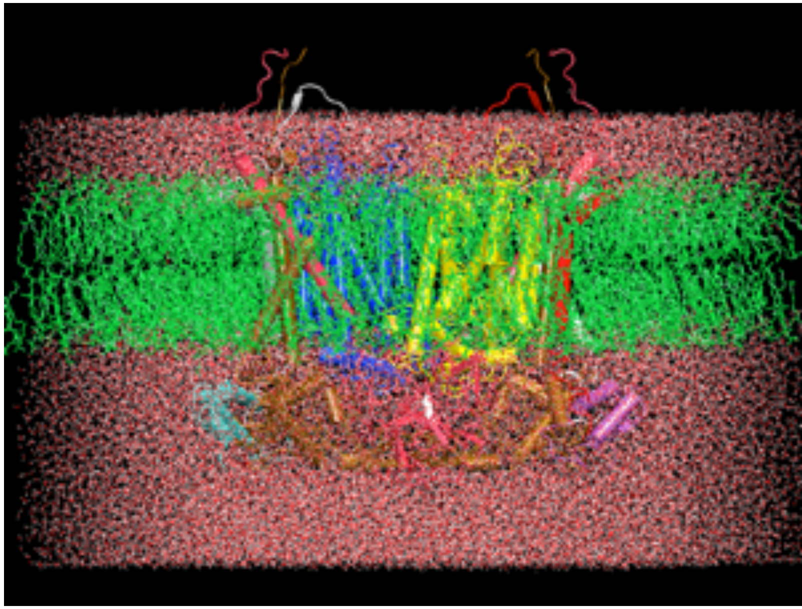
Steps in a Typical MD Simulation

- **1. Prepare molecule**
  - Read in pdb and psf file
- **2. Minimization**
  - Reconcile observed structure with force field used ($T = 0$)
- **3. Heating**
  - Raise temperature of the system
- **4. Equilibration**
  - Ensure system is stable
- **5. Dynamics**
  - Simulate under desired conditions (NVE, NpT, etc)
  - Collect your data
- **6. Analysis**
  - Evaluate observables (macroscopic level properties)
  - Or relate to single molecule experiments



Preparing Your System for MD *solvation*

Biological activity is the result of interactions between molecules and occurs at the interfaces between molecules (protein-protein, protein-DNA, protein-solvent, DNA-solvent, etc).



Why model solvation?

• many biological processes occur in aqueous solution

• solvation effects play a crucial role in determining molecular conformation, electronic properties, binding energies, etc.

How to model solvation?

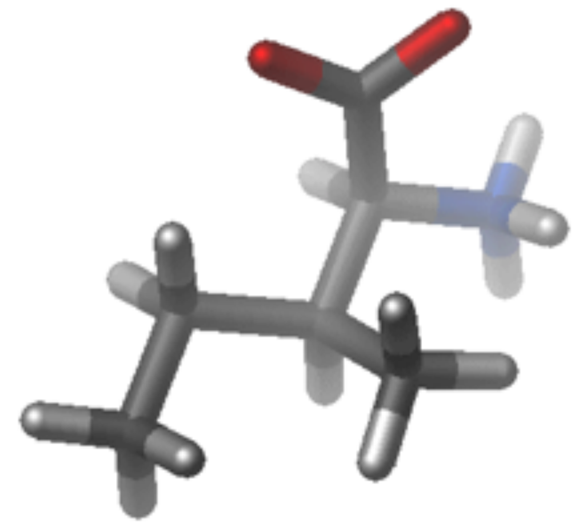
• explicit treatment: solvent molecules are added to the molecular system

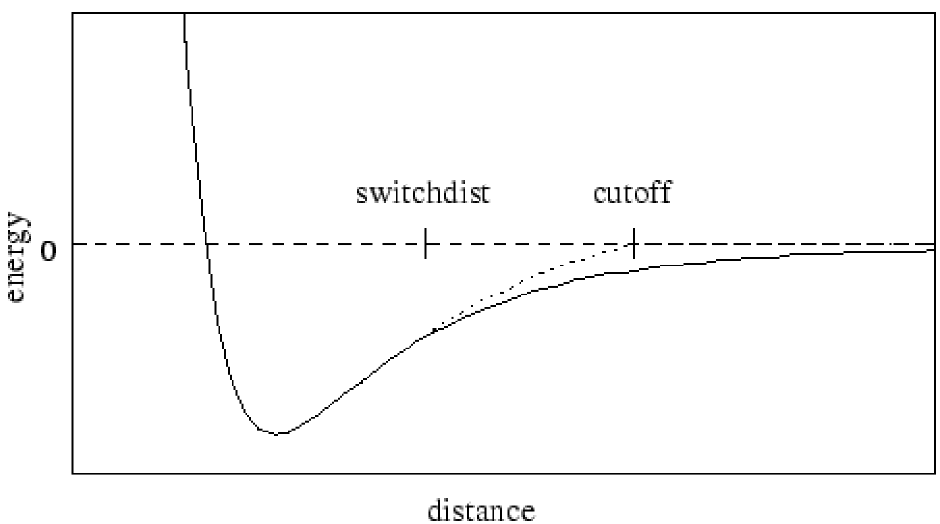
• implicit treatment: solvent is modeled as a continuum dielectric with additional forces (**Generalized Born Implicit Solvent** in NAMD)

Computational tricks are needed

-bonded terms easy to calculate,
can be done every time step
-vdW, Coulomb terms scale as N^2 however,
need approximations to efficiently calculate

-***exclude scaled1-4*** eliminates non-bonded
interactions between atoms 1-2 bonds apart,
scales those 3 apart (*these non-bonded
interactions are already in the bonded terms*)





-cutoff tail of vdW potential

-strict cutoff won't conserve energy,
introduces artifacts

-instead change the function to
smoothly approach zero

Multiple time stepping (MTS or R-RESPA)

-Interactions at different distances vary on different timescales

-NAMD defines three regimes - bonded interactions, short-range non-bonded, and long-range electrostatics

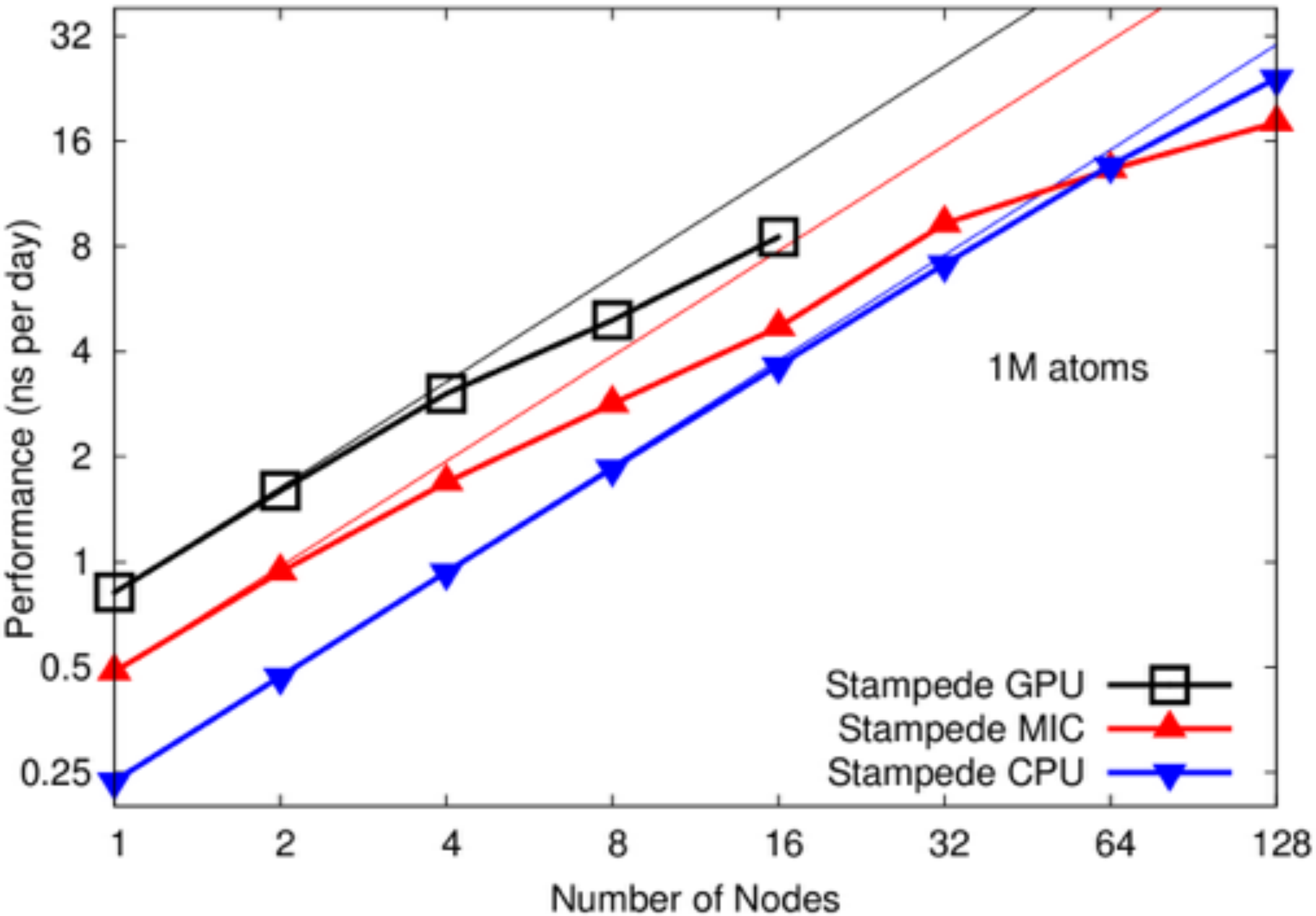
-bonded are calculated every time step; non-bonded (vdW and short range electrostatics) can be calculated less often; full electrostatics even less frequently (using PME - particle mesh Ewald - $N \log N$ scaling)

-**MTS** schemes greater than 1-1-3 or 2-1-2 not recommended due to resonances that can occur (1-2-4 or 2-1-3 in case of NVT/NPT)

-**MTS** also introduces energy, and therefore temperature, drift in NVE ensemble; can be overcome by using a thermostat (i.e., NVT/NPT ensemble)

NAMD: The program we will use

NAMD 2.11 on STMV (2-fs time step, 12A cutoff + PME every 3 steps) on TACC Stampede



NAMD programmer
Jim Phillips
PhD UIUC Physics

Simulation of large biomolecular systems

2002 Gordon Bell Award for parallel scalability.

Runs at NSF centers, on clusters, and on desktop.

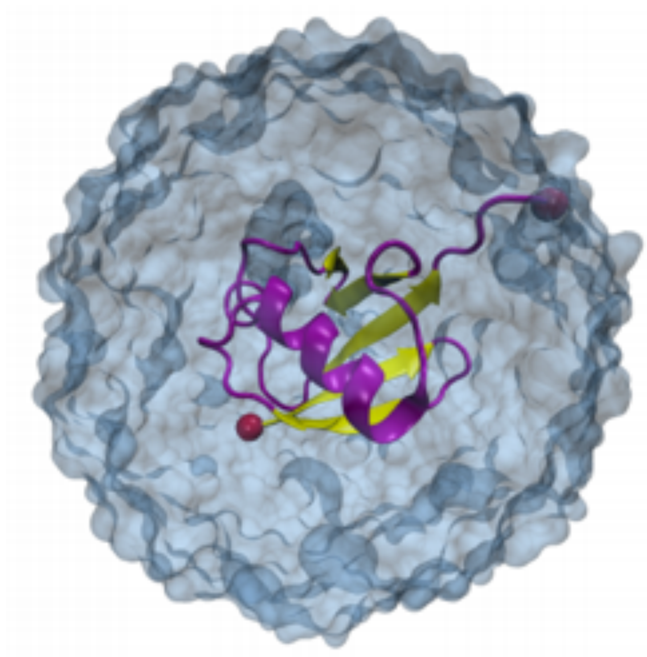
Available for **FREE** as precompiled binaries; includes source code.

10,000+ registered users.

NAMD
Scalable Molecular Dynamics

NAMD TUTORIAL

Unix/MacOSX Version



NAMD Developers: James Phillips, David Hardy

NAMD Tutorial Contributors: Tim Isgro, James Phillips, Marcos Sotomayor, Elizabeth Villa, Hang Yu, David Tanner, Yanxin Liu, Zhe Wu, David Hardy

October 2016

The NAMD Configuration File / 1

Files needed:

```
structure          mypsf.psf
coordinates        mypdb.pdb
```

Define temperature

```
set temperature    310
     ;# target temperature used several times below
```

Starting simulation with random velocities

```
# starting from scratch
temperature            $temperature
      ;# initialize velocities randomly
```

The NAMD Configuration File / 2

Continuing a simulation with positions and velocities from previous run

```
# continuing a run
set inputname       myinput           ;# only need to edit this in one place!
binCoordinates      $inputname.coor   ;# coordinates from last run (binary)
binVelocities       $inputname.vel    ;# velocities from last run (binary)
extendedSystem     $inputname.xsc    ;# cell dimensions from last run
firsttimestep       50000             ;# last step of previous run
numsteps            100000            ;# run stops when this step is reached
```

The NAMD Configuration File / 3

Organizing output

```
outputName          myoutput
        ;# base name for output from this run

restartfreq         500     ;# 500 steps = every 1ps
dcdfreq             500
xstFreq             500

outputEnergies     100     ;# 100 steps = every 0.2 ps
outputTiming      1000
        ;# shows time per step and time to completion
```

The NAMD Configuration File / 4

```
# Force-Field Parameters
paraTypeCharmm     on
parameters          par_all27_prot_lipid.inp

# These are specified by CHARMM
exclude             scaled1-4
1-4scaling          1.0
switching           on

# You have some freedom choosing the cutoff
cutoff              12. ;#
switchdist          10. ;# cutoff - 2.

# Promise that atom won't move more than 2A in a cycle
pairlistdist        14. ;# cutoff + 2.
stepspercycle       10   ;# redo pairlists every ten steps
```

Energy drifts if too large, but smaller requires more steps per ns.

```
# Integrator Parameters
timestep            2.0  ;# 2fs/step
rigidBonds          all  ;# needed for 2fs steps
nonbondedFreq       1    ;# nonbonded forces every step
fullElectFrequency  2    ;# PME only every other step
```

The NAMD Configuration File / 5

Controlling temperature

```
# Constant Temperature Control
langevin            on              ;# langevin dynamics

langevinDamping     1               ;# damping coefficient of 1/ps
langevinTemp        $temperature    ;# random noise at this level
langevinHydrogen    no              ;# don't couple bath to hydrogens
```

Underlying Langevin equation for all atoms

$$m_i \frac{d^2 x_i(t)}{dt^2} = F_{i,\mathrm{ff}} - \gamma m_i \frac{d x_i(t)}{dt} + R_i(t)$$

The NAMD Configuration File / 6

Using periodic boundary conditions
avoids surface effects; permits particle mesh Ewald (PME) electrostatics; permits pressure control

```
# Periodic Boundary conditions
cellBasisVector1    31.2   0.    0.   ;# vector to the next image
cellBasisVector2     0.   44.8  0.
cellBasisVector3     0.     0   51.3
cellOrigin           0.     0.    0.   ;# the *center* of the cell

wrapWater           on                 ;# wrap water to central cell
wrapAll             on                 ;# wrap other molecules too
wrapNearest         off                ;# use for non-rectangular cells
```

The NAMD Configuration File / 7

particle mesh Ewald electrostatics
(avoids cut-off of long-range Coulomb forces)

```
#PME (for full-system periodic electrostatics)
PME                   yes
PMEGridSizeX          32  ;# 2^5, close to 31.2
PMEGridSizeY          45  ;# 3^2 * 5, close to 44.8
PMEGridSizeZ          54  ;# 2 * 3^3, close to 51.3
```

Typically, it’s easier to just let NAMD choose the PME grid parameters

```
# PME (for full-system periodic electrostatics)
PME                   yes
PMEGridSpacing        1.0
```

The NAMD Configuration File / 8

Fix atoms

```
fixedAtoms          on
fixedAtomsFile      myfixedatoms.pdb  ;# flags are in this file
fixedAtomsCol       B                 ;# set beta non-zero to fix an atom
```

The NAMD Configuration File / 9

Energy-minimize structure (*T*=0) , reset temperature *T*, run:

```
minimize            1000            ;# lower potential energy for 1000 steps
reinitvels          $temperature    ;# since minimization zeros velocities
run 50000 ;# 100ps
```

The NAMD Output File / 1

Preamble

```
Info: NAMD 2.5b2ss03 for Linux-i686-Clustermatic
Info:
Info: Please visit http://www.ks.uiuc.edu/Research/namd/
Info: and send feedback or bug reports to namd@ks.uiuc.edu
Info:
Info: Please cite Kale et al., J. Comp. Phys. 151:283-312 (1999)
Info: in all publications reporting results obtained with NAMD.
Info:
Info: Built Fri May 30 13:09:06 CDT 2003 by jim on umbriel
Info: Sending usage information to NAMD developers via UDP.
Info: Sent data is: 1 NAMD  2.5b2ss03  Linux-i686-Clustermatic 47 umbriel jim
Info: Running on 47 processors.
```

The NAMD Output File / 2

Energies

```
ETITLE:      TS           BOND          ANGLE          DIHED          IMPRP
          ELECT            VDW       BOUNDARY           MISC        KINETIC
          TOTAL           TEMP         TOTAL2         TOTAL3        TEMPAVG
       PRESSURE      GPRESSURE         VOLUME      PRESSAVG      GPRESSAVG

ENERGY:    1000         0.0000         0.0000         0.0000         0.0000
    -97022.1848      9595.3175         0.0000         0.0000     14319.5268
    -73107.3405       300.2464    -73076.6148    -73084.1411       297.7598
      -626.5205      -636.6638    240716.1374      -616.5673      -616.6619
```

The NAMD Output File / 3

Writing out trajectories

⋮

```
OPENING COORDINATE DCD FILE
WRITING COORDINATES TO DCD FILE AT STEP 1000
```

⋮

Performance information

```
Info: Benchmark time: 47 CPUs 0.0475851 s/step 0.275377 days/ns 13540 kB memory

TIMING: 1000  CPU: 18.35, 0.01831/step  Wall: 50.1581, 0.0499508/step, 6.92374
hours remaining, 14244 kB of memory in use.
```

Warnings

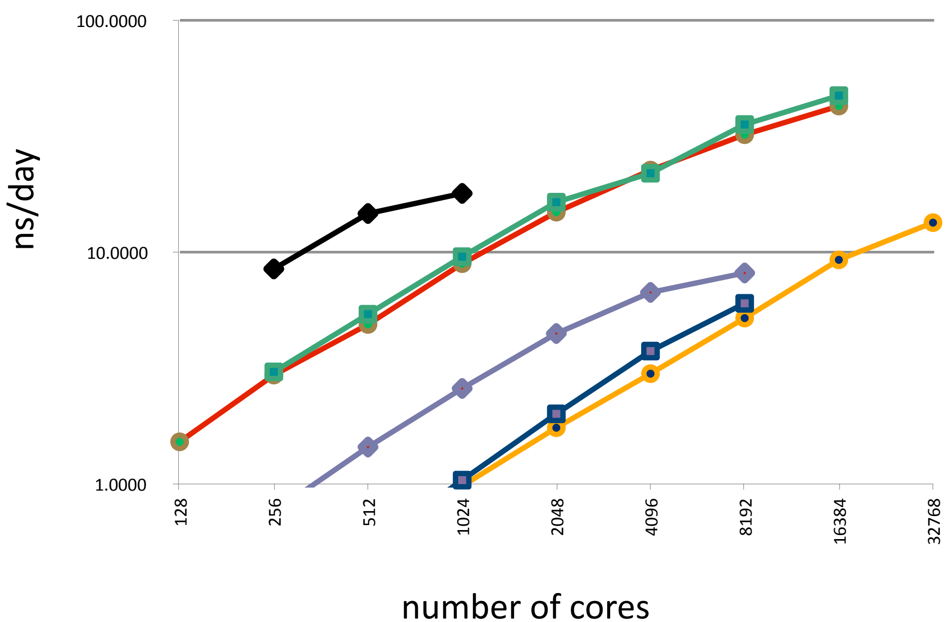
```
Warning: Pairlistdist is too small for 1 patches during timestep 17.
Warning: Pairlists partially disabled; reduced performance likely.
Warning: 20 pairlist warnings since previous energy output.
```

Measuring performance

Check your scaling!!!

grep "Benchmark" *log

```
Info: Benchmark time: 42 CPUs 0.0879267 s/step 1.01767 days/ns 87.665 MB memory
```



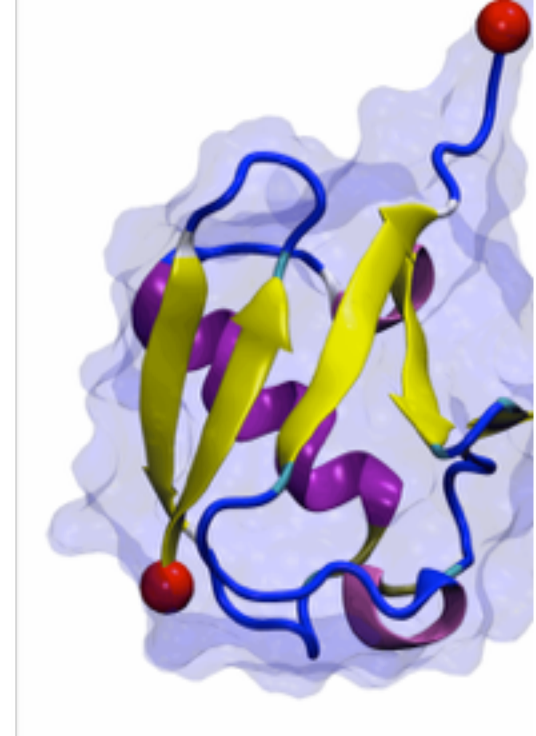
$$\textbf{Efficiency} = \frac{\#\ \text{s/step (1 cpu)}}{n * \#\ \text{s/step } (n \text{ cpus})}$$

TYPICAL RANGE:
250-1000 atoms/core

grep "TIMING" *log

```
TIMING: 3000  CPU: 346.34, 0.07938/step  Wall: 466.648, 0.0879514/step,
6.08331 hours remaining, 88.341812 MB of memory in use.
```

Dealing with crashes



Some errors are obvious...

“Cannot specify both an initial temperature and a velocity file”

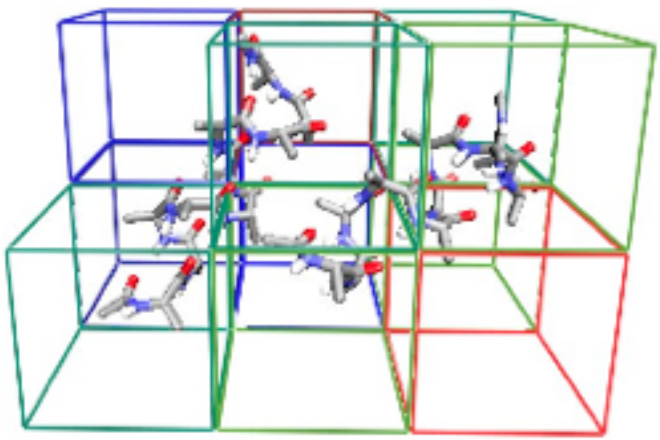
"stepsPerCycle must be a multiple of fullElectFrequency"

etc...

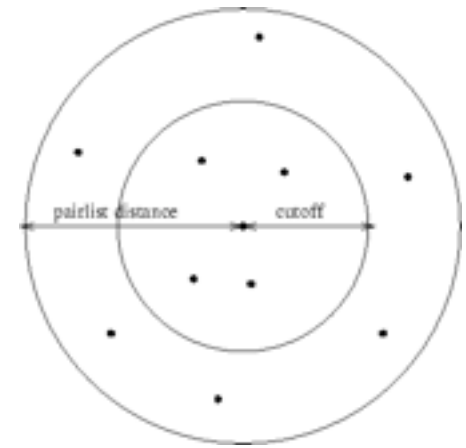
Others not so much...

```
FATAL ERROR: Periodic cell has become too small for original patch grid!
Possible solutions are to restart from a recent checkpoint,
increase margin, or disable useFlexibleCell for liquid simulation.
```

-relates to how NAMD parallelizes the simulation



Atoms that move close enough to interact (defined by cutoff) but are not on neighboring patches causes a crash



-typically happens because of large volume fluctuations (normal during initial equilibriation in NpT ensemble), but **CHECK OUTPUT TO BE SURE**

**can set “margin 2” (force bigger patches) in configuration file
**lower “stepsPerCycle” in configuration file
**just restart

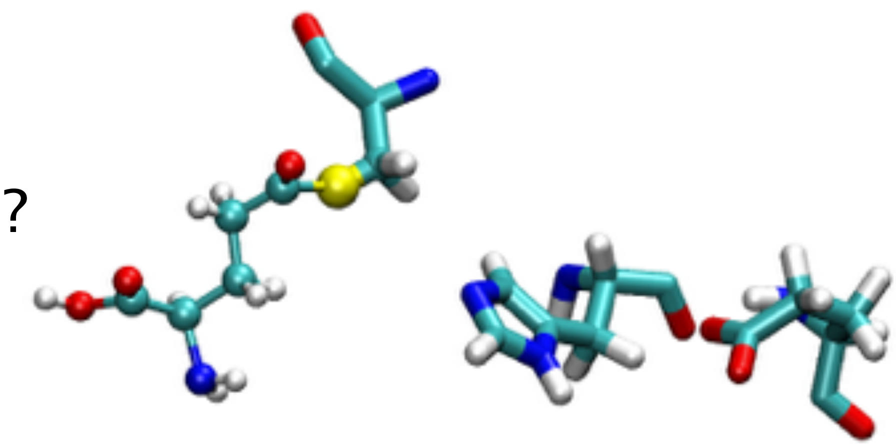
Missing parameters?

```
DIDN'T FIND vdW PARAMETER FOR ATOM TYPE CT3
```

**Did you specify all the needed parameter files?

**Was your system (PSF/PDB) constructed correctly? (Check for errors/warnings from PSFGen or AutoPSF!)

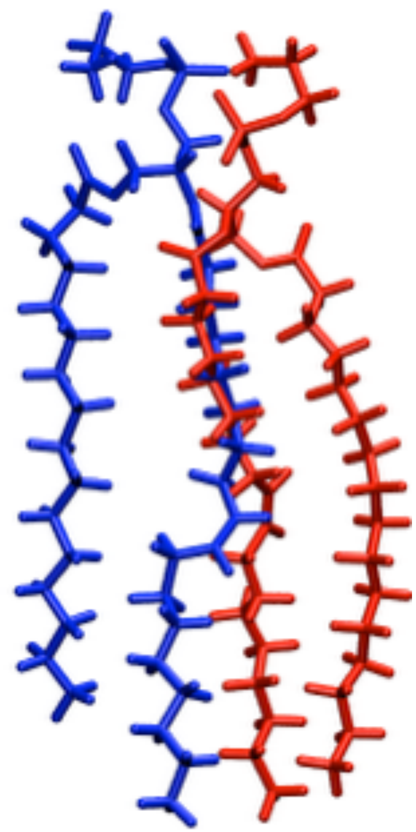
**Do you have an unusual ligand? (need to either remove or develop parameters for it)



Simulation instability

```
ERROR: Atoms moving too fast; simulation has become unstable.
ERROR: Constraint failure in RATTLE algorithm for atom 1897!
```

Both errors almost always derive from bad system configurations!

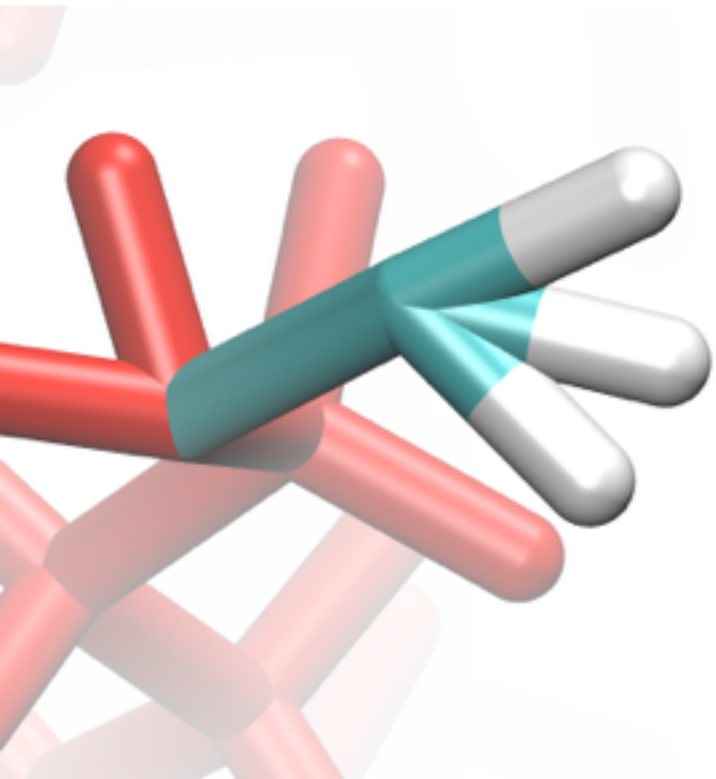


**Check your system in VMD near the noted atoms
**Use the "measure contacts" command to check for atoms that are very close (say, within 0.1 Å)
**Look for atoms at (0,0,0) whose positions didn't get initialized when building PSF/PDB
**Check that the periodic box dimensions are big enough
**Minimize for longer, or set margin higher
If all else fails, change your DCDFreq to 1 and watch the simulation up to the point of the crash **very carefully**

Simulation instability (cont.)

```
FATAL ERROR: Bad global exclusion count.
```

typically results from bad starting configuration, similar to previous errors



-besides previous solutions, consider the possibility of missing angle or dihedral entries from the PSF file

-when a patch is applied by PSFGen, the command
"**regenerate angles dihedrals**"
may need to be issued before guessing coordinates and writing the PSF

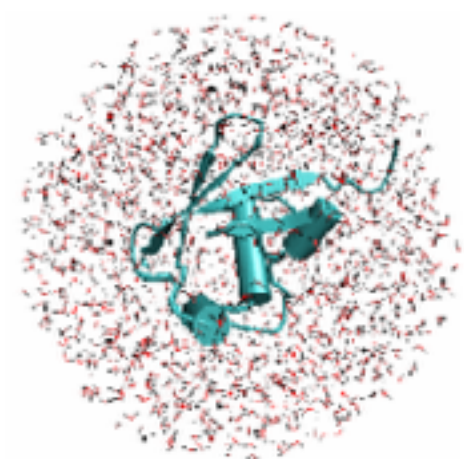
The NAMD Experience/ 1

You will first simulate ubiquitin in a water sphere and water box:

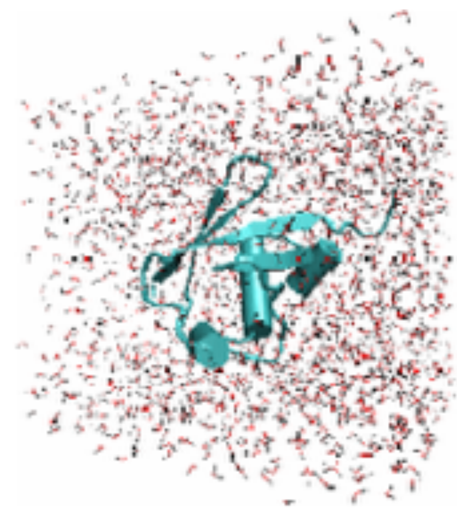
Generating a Protein Structure File (PSF)

- Go to 1-1-build directory
- Open VMD, choose extension TkCon
- Make from 1UBQ.pdb a structure without hydrogens, ubqp.pdb
- Create psf file for ubqp.pdb: ubq.pdb and ubq.psf
- Check if files exist

Solvate the protein in a water sphere (from VMD)



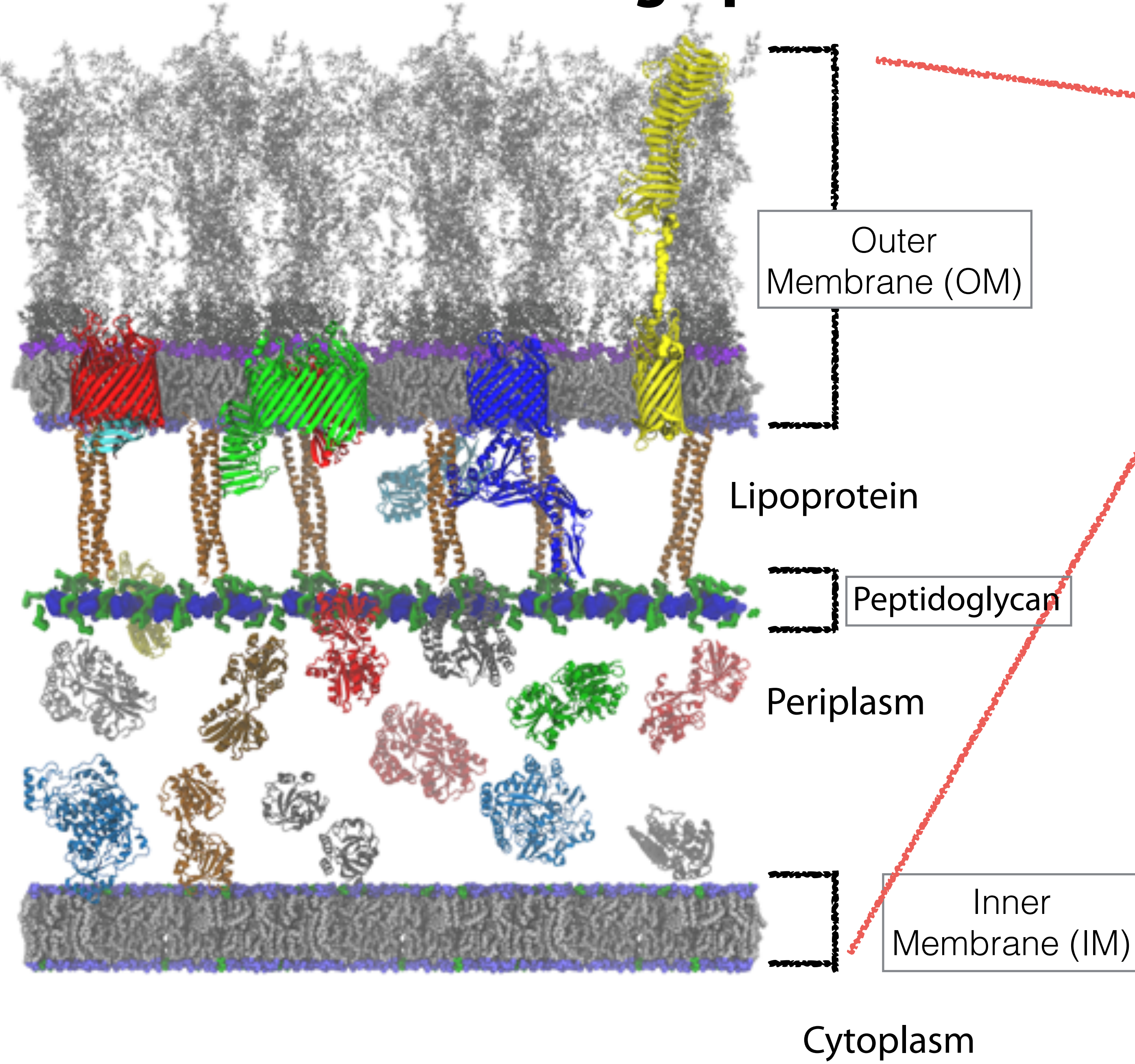
Solvate the protein in a water box (from VMD)

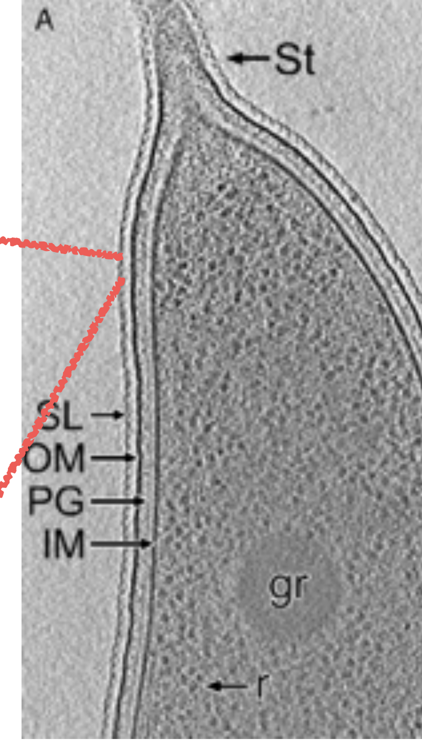


The NAMD Experience/ 2

- RMSD value for equilibration
- Atomic RMSD values of equilibrated protein
- Velocity distribution
- Temperature distribution
- Specific heat
- Diffusion of whole protein
- Heat diffusion
- Temperature Echoes

cracking open bacteria



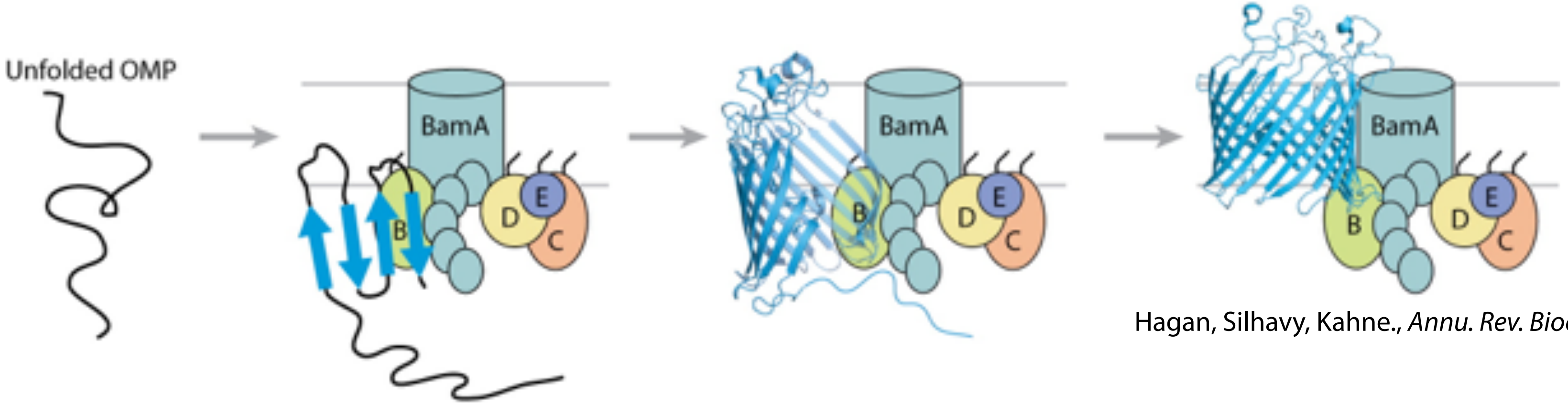


From Gan, Chen, Jensen. (2008) *PNAS* **105**:18953-57.

Gram negative bacteria cell envelope:
two layers of membrane and a rigid cell wall

Escherichia coli

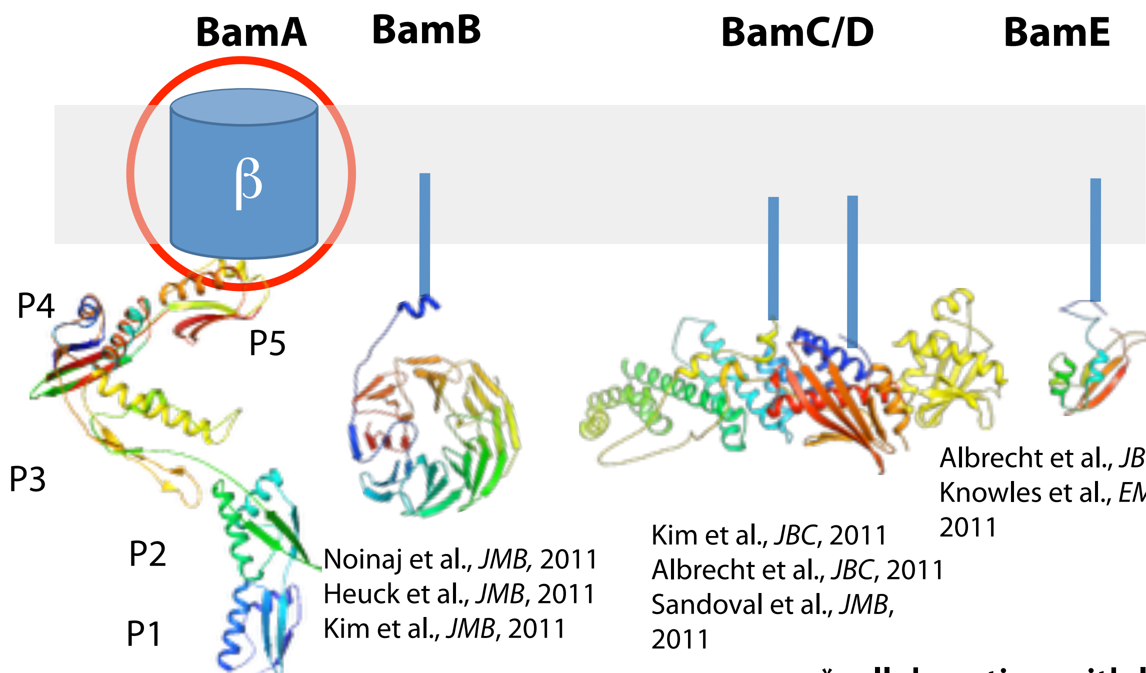
Membrane insertion *at the outer membrane*



Hagan, Silhavy, Kahne., *Annu. Rev. Biochem.*, 2011.

no energy source at the OM - β-barrels require a novel way to fold!

Known structures of Bam complex proteins



Gatzeva-Topalova et al., *Structure*, 2008
Kim et al., *Science*, 2007
Gatzeva-Topalova et al., *Structure*, 2010

Noinaj et al., *JMB*, 2011
Heuck et al., *JMB*, 2011
Kim et al., *JMB*, 2011

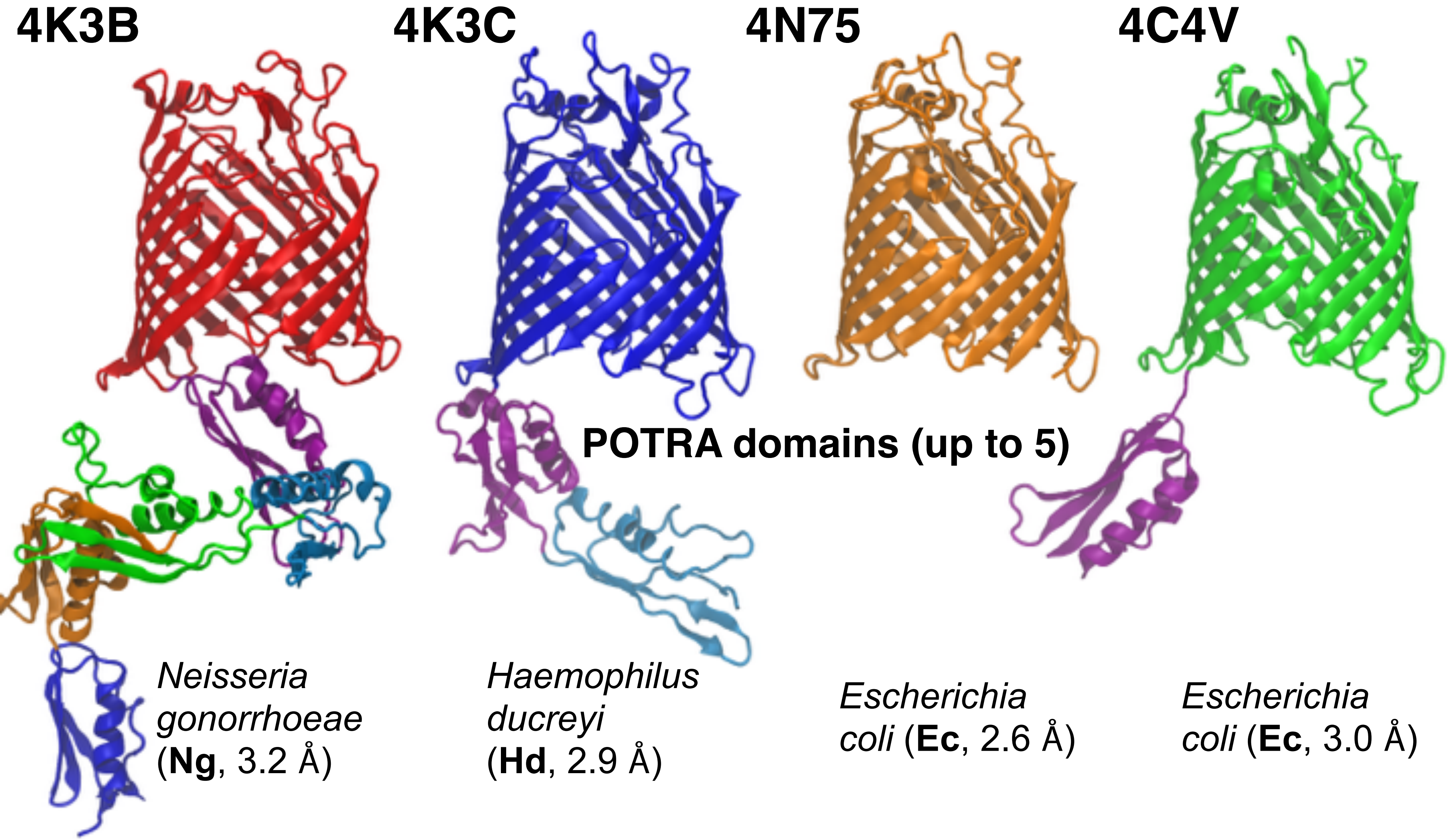
Kim et al., *JBC*, 2011
Albrecht et al., *JBC*, 2011
Sandoval et al., *JMB*, 2011

Albrecht et al., *JBC*, 2011
Knowles et al., *EMBO Rep*, 2011

BamA structure was the last to be solved

***collaboration with lab of Susan Buchanan (NIH) and Nick Noinaj (Purdue)**

The missing piece - crystal structures of BamA

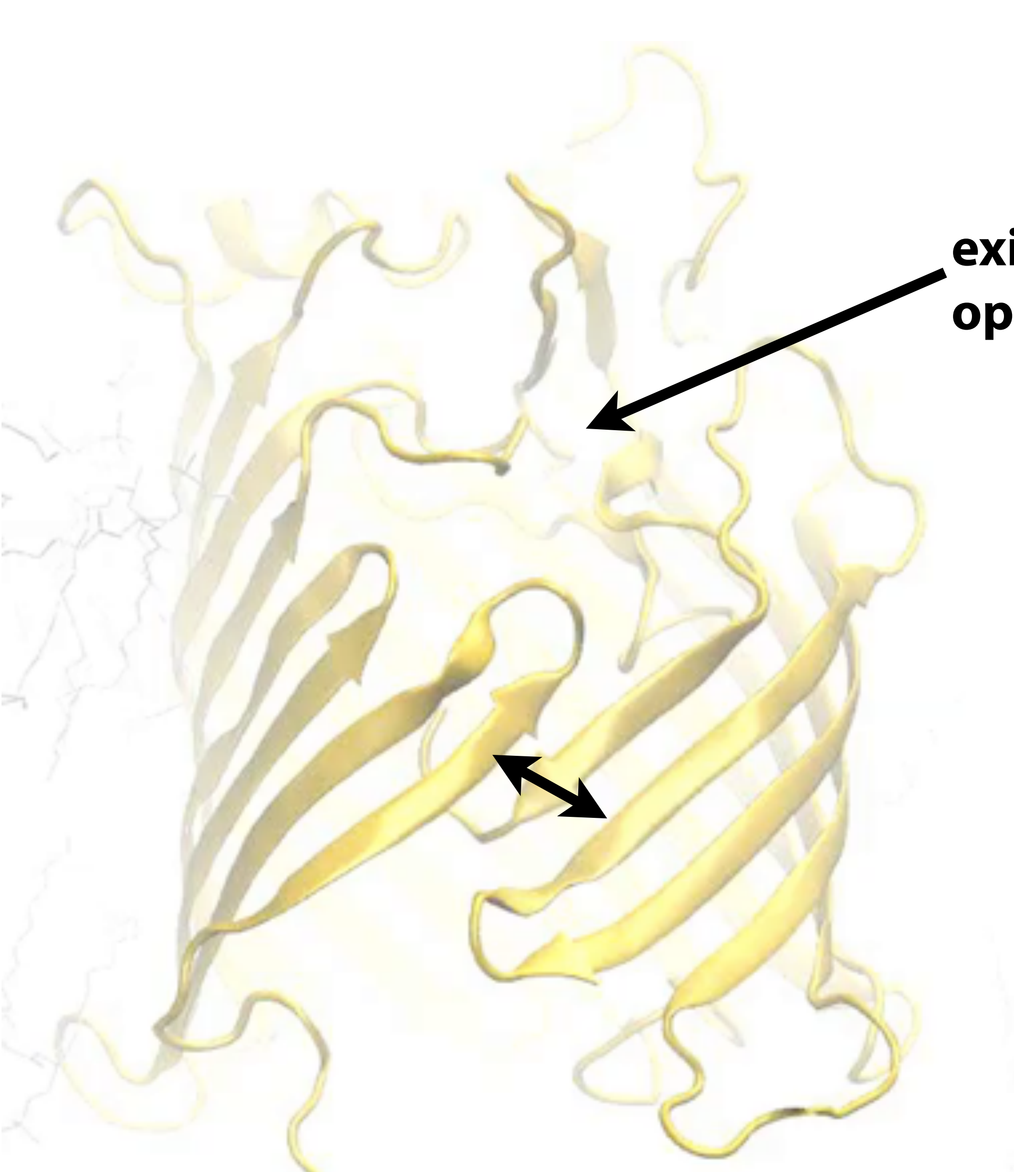


N. Noinaj…S. K. Buchanan. (2013) *Nature*. 501:385-390.

D. Ni…Y. Huang. (2014) *FASEB J*. 28:2677-2685.

R. Albrecht…K. Zeth. (2014) *Acta Cryst*. D70:1779-1789.

µs-time-scale simulation of NgBamA



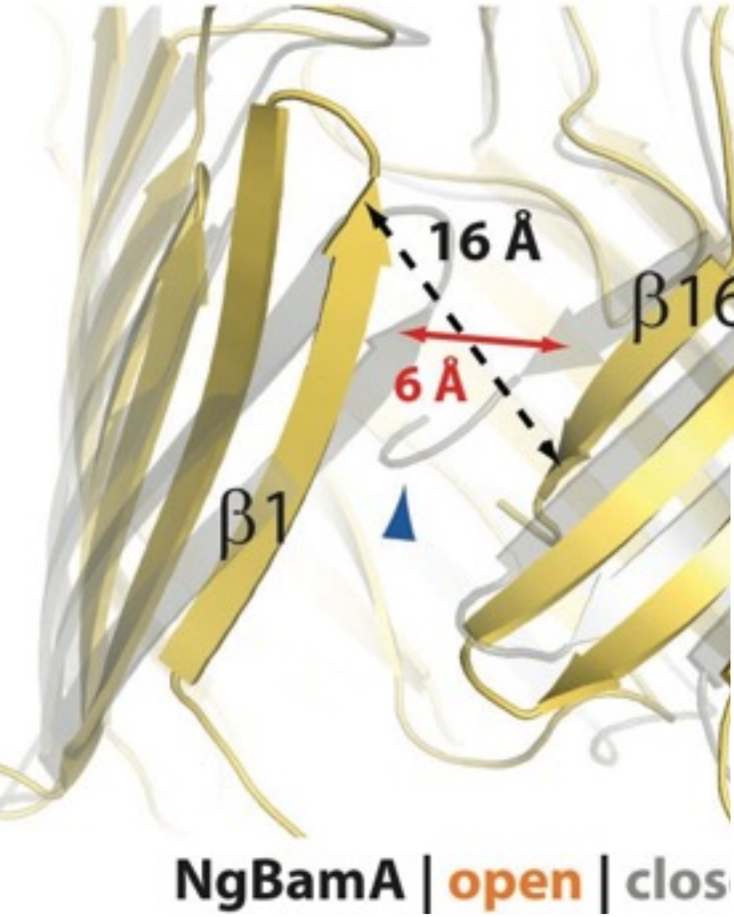
exit pore at the top also opens up

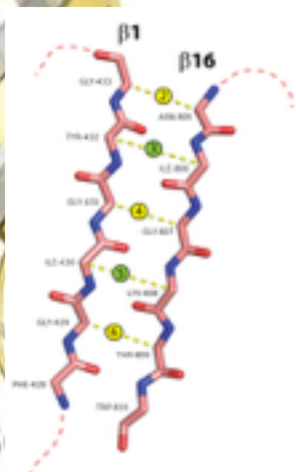
“lateral gate” between N/C barrel strands opens (and closes) spontaneously!

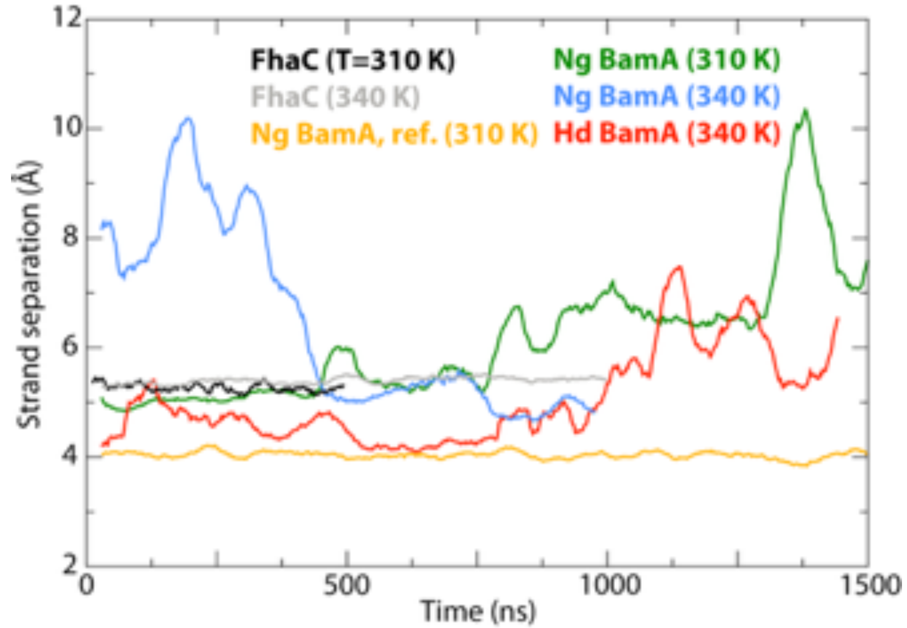
N. Noinaj, A. J. Kuszak, C. Balusek, J.C. Gumbart, and S. K. Buchanan. (2014) *Structure.* 22:1055-1062.

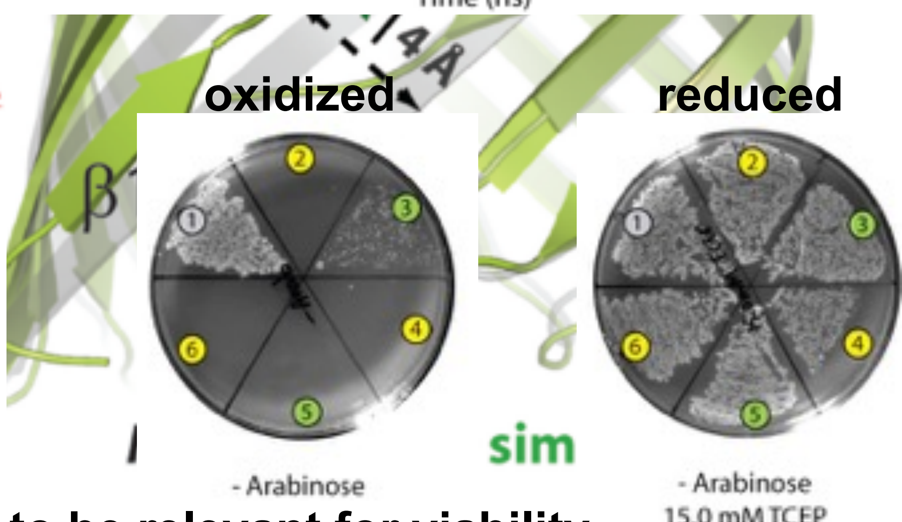
N. Noinaj, A. J. Kuszak, J.C. Gumbart, P. Lukacik, H. Chang, N. Easley, T. Lithgow, and S. K. Buchanan. (2013) *Nature*. 501:385-390.

Simulations reveal dynamic gating









gating confirmed to be relevant for viability

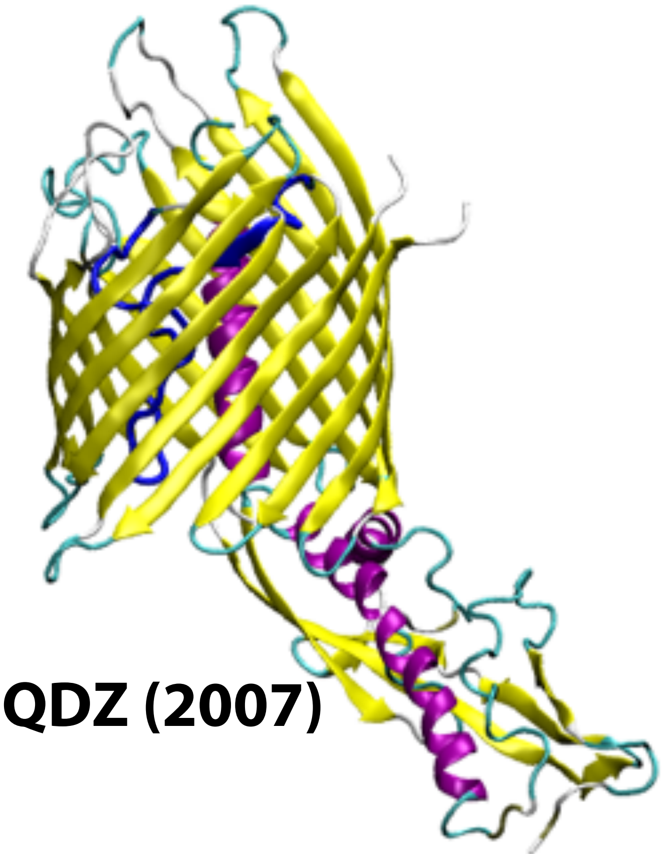
N. Noinaj…S. K. Buchanan. (2013) *Nature*. 501:385-390.

N. Noinaj…S. K. Buchanan. (2014) *Structure*. 22:1055-1062.

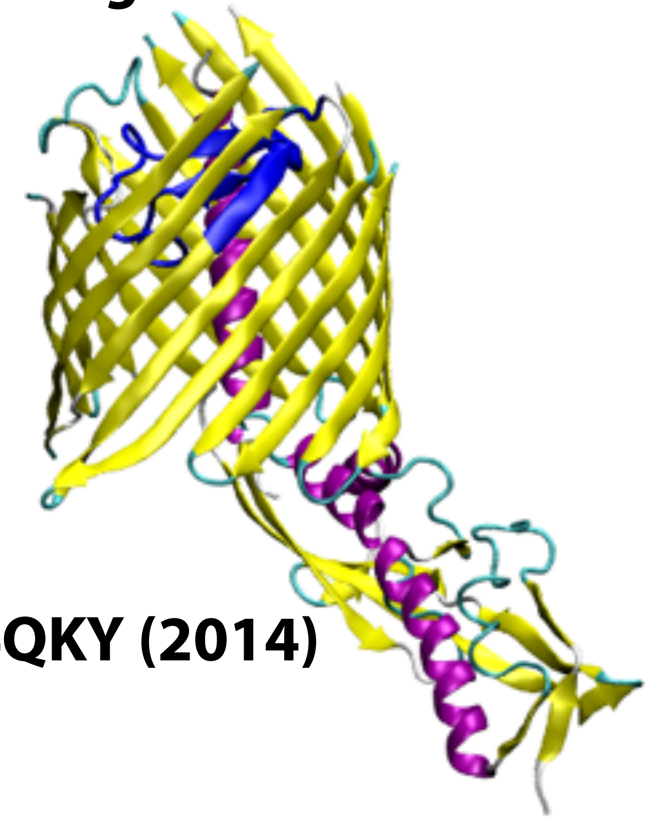
The cautionary tale of FhaC

FhaC is a homologous protein to BamA, but no lateral gate

register shift > 20 residues!



2QDZ (2007)



4QKY (2014)

2QDZ was obsoleted on 2014-10-22 and superseded by 4QKY

new revised structure from the **same** diffraction data!