MD Simulations of Proteins: Practical Hints and Pitfalls to Avoid

Stefan Boresch
stefan@mdy.univie.ac.at

Department of Computational Biological Chemistry
Faculty of Chemistry, University of Vienna

Vienna Summer School on Drug Design, September 2017

Molecular dynamics (MD) in a nutshell

One particle

force=mass×acceleration

$$\mathbf{F} = m\,\mathbf{a}$$

i.e.

$$\frac{d^2\mathbf{r}}{dt^2} = \ddot{\mathbf{r}} = \frac{1}{m}\mathbf{F}$$

The position $\mathbf{r}(t)$ of the particle is described by a 2nd order differential equation (Initial condition: $\mathbf{r}$ and $\mathbf{v}$ at $t = 0$)

N particles

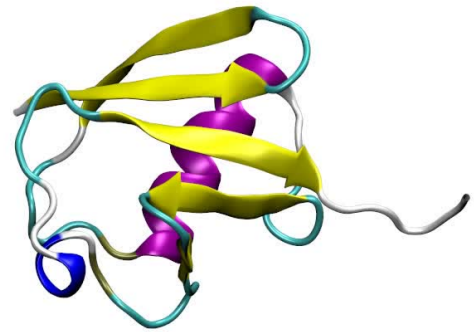
$$\ddot{\mathbf{r}}_1 = \frac{1}{m_1}\mathbf{F}_1(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$$

$$\ddot{\mathbf{r}}_2 = \frac{1}{m_1}\mathbf{F}_2(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$$

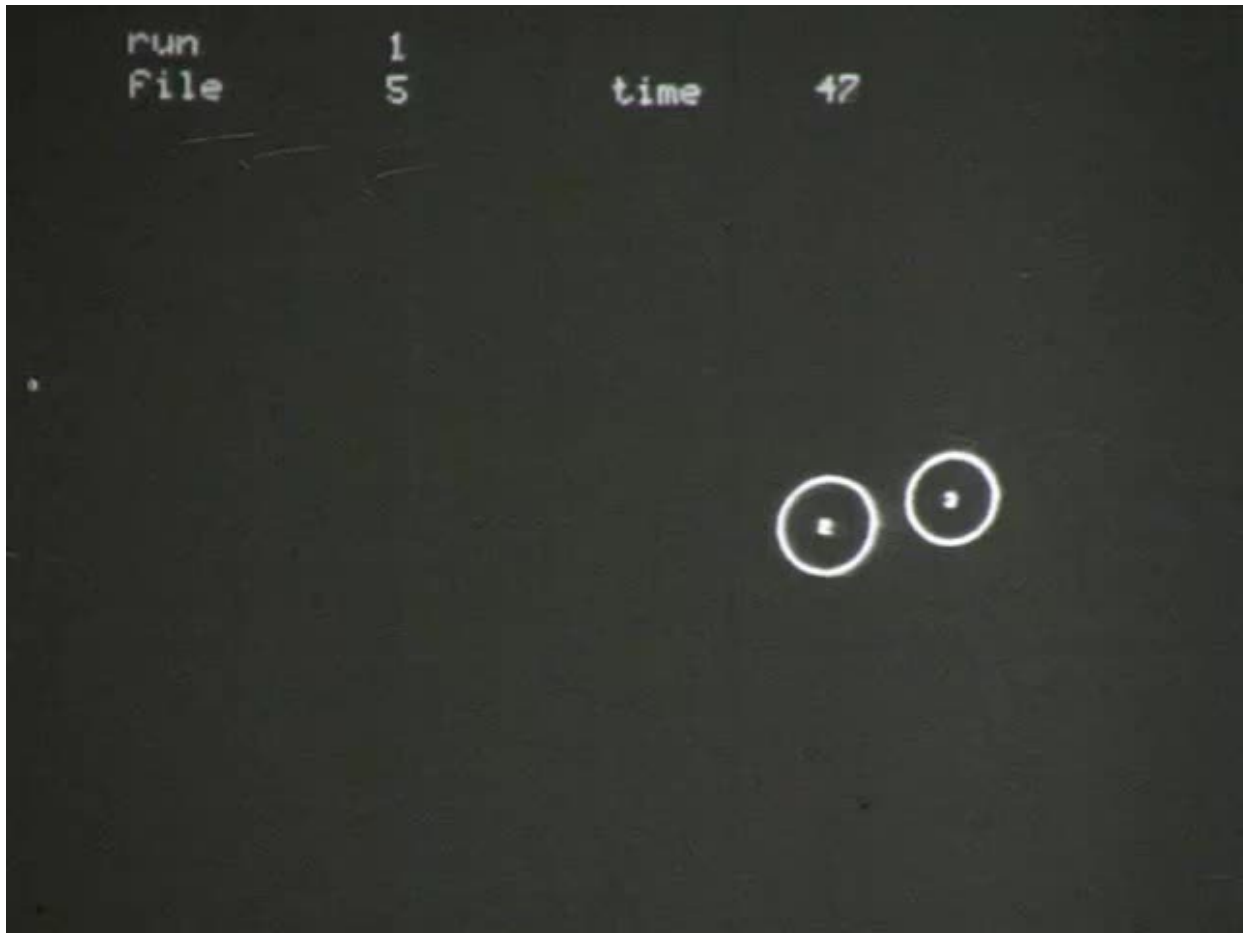
$$\ldots$$

$$\ddot{\mathbf{r}}_N = \frac{1}{m_N}\mathbf{F}_N(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$$

Numerical integration:



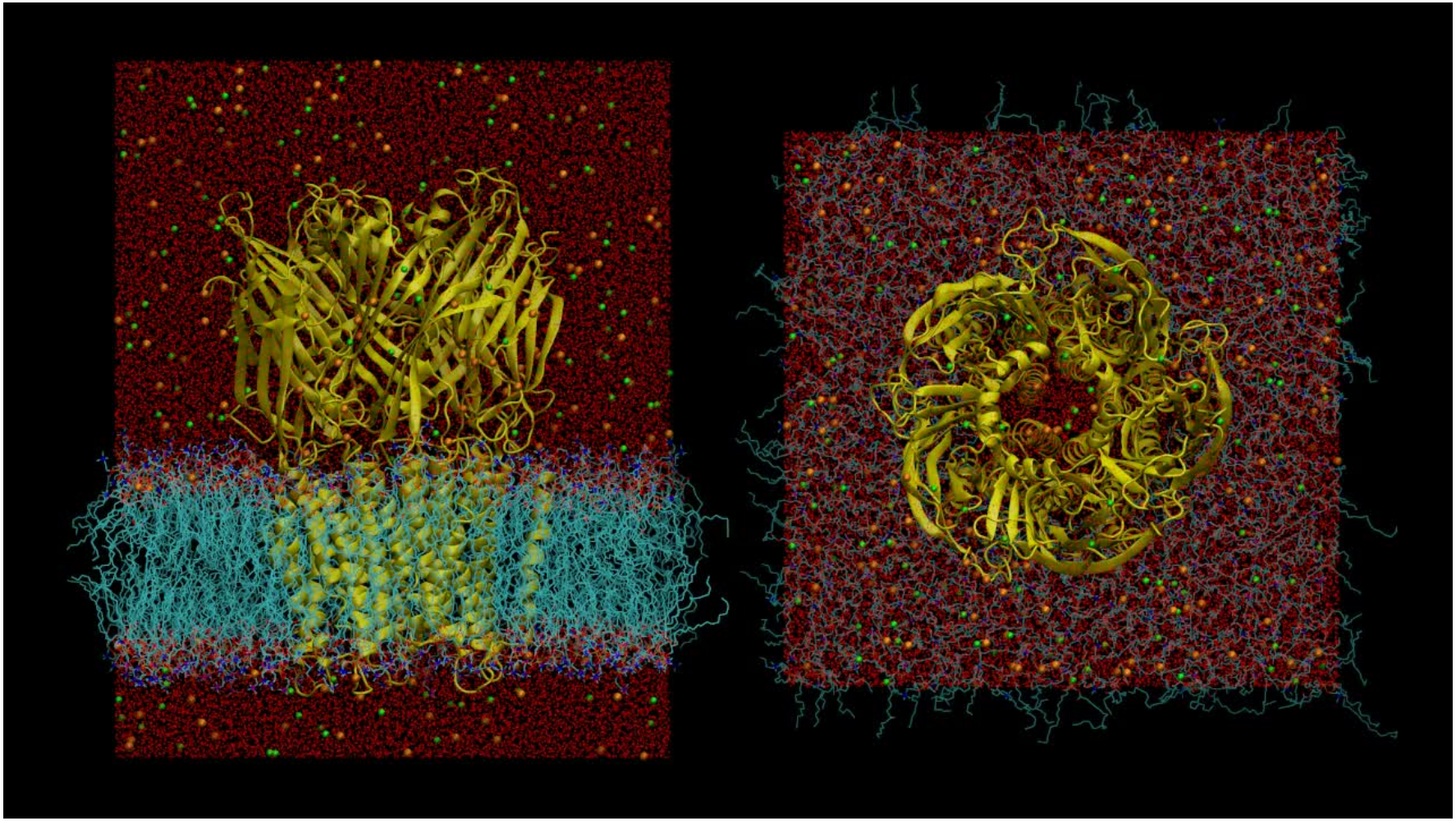
How it all began — the $H + H_2$ reaction*



(Martin Karplus & co-workers, 1964-67)

*Movie/image courtesy of Martin Karplus

50 years later . . . †



(PNAS 2013, 110, E3987 doi: 10.1073/pnas.1313785110)

† Movie/image courtesy of Marco Cecchini

Ingredients of MD simulations of biomolecular systems

- Approximate, fast description of interactions
- Numerical integration of the (classical) equations of motion
- Analysis of the data (i.e., trajectories)

Ingredients of MD simulations of biomolecular systems

- Approximate, fast description of interactions
  $\Rightarrow$ force fields
- Numerical integration of the (classical) equations of motion
  $\Rightarrow$ several excellent programs available
- Analysis of the data (i.e., trajectories)

Ingredients of MD simulations of biomolecular systems

- Approximate, fast description of interactions
  $\Rightarrow$ force fields
- Numerical integration of the (classical) equations of motion
  $\Rightarrow$ several excellent programs available
- Analysis of the data (i.e., trajectories)
  $\Rightarrow$ still the "final frontier"

Ingredients of MD simulations of biomolecular systems

- Approximate, fast description of interactions
  $\Rightarrow$ force fields
- Numerical integration of the (classical) equations of motion
  $\Rightarrow$ several excellent programs available
- Analysis of the data (i.e., trajectories)
  $\Rightarrow$ still the "final frontier"
- Assembling and setting up the simulation system
  - Build a meaningful model of the real problem
  - Validate experimental data and/or fill in missing pieces
  - Make any necessary "educated guesses" concerning the state of your system

Concerning analysis

Think about it before you start simulating: *What quantities do you want to compute — how are you going to extract them from the simulation "raw data" (i.e., trajectories)*?

Monitor your system (routine analysis) throughout all stages of the project!

Force fields

Pairwise, additive force fields: Eq. 27 of Lifson and Warshel, JCP 49, 5116 (1968)

of $\partial V(\mathbf{r}, \mathbf{x}+\delta\mathbf{x}_m)/\partial r_\alpha = 0$ by Eq. (5),

$$\mathbf{r}_0(\mathbf{x}+\delta\mathbf{x}_m) = \mathbf{r}(\mathbf{x}+\delta\mathbf{x}_m) - F^{-1}(\mathbf{r}, \mathbf{x}+\delta\mathbf{x}_m)\nabla V(\mathbf{r}, \mathbf{x}+\delta\mathbf{x}_m). \quad (21)$$

It is possible to choose $\mathbf{r}(\mathbf{x}+\delta\mathbf{x}_m)$ such that

$$\mathbf{r}(\mathbf{x}+\delta\mathbf{x}_m) = \mathbf{r}_0(\mathbf{x}). \quad (22)$$

From Eqs. (20)–(22), it follows that

$$\frac{\partial \mathbf{r}_0}{\partial x_m} = \lim_{\delta x_m \to 0} \frac{-\mathbf{F}^{-1}[\mathbf{r}_0(\mathbf{x}); \mathbf{x}+\delta\mathbf{x}_m]\nabla V[\mathbf{r}_0(\mathbf{x}); \mathbf{x}+\delta\mathbf{x}_m]}{\delta x_m} = -\mathbf{F}^{-1}(\mathbf{r}_0; \mathbf{x})\partial\nabla V(\mathbf{r}_0; \mathbf{x})/\partial x_m. \quad (23)$$

The expressions $\mathbf{F}^{-1}$ and $\mathbf{F}$ have been used already in the derivation of $\mathbf{r}_0$ and ν_α, and $\partial\nabla V/\partial x_m$ is derived from analytical expressions of the gradient ∇V as explicit function of $\mathbf{x}$.

(3) Overend and Scherer[17] considered the normal-mode frequencies as functions of the force constants and used the least-squares method to obtain optimal frequencies, equilibrium conformations, conformational strain energies, and enthalpies of vibration–rotation–translation. The method was first tested in the set of functions used by Bixon and Lifson,[4] with a few modifications, to allow the H atoms to participate in the vibrational modes. With this set of functions the molecular energy is given by

$$V(\mathbf{s}) = \tfrac{1}{2}\sum_i K_b(b_i-b_0)^2 + \tfrac{1}{2}\sum_i K_\theta(\theta_i-\theta_0)^2 + \tfrac{1}{2}\sum_{i,\sigma} K_\alpha(\alpha_i^\sigma-\alpha_0)^2 + \tfrac{1}{2}\sum_{i,\sigma} K_\gamma(\gamma_i^\sigma-\gamma_0)^2 + \tfrac{1}{2}\sum_i K_\delta(\delta_i-\delta_0)^2 + \tfrac{1}{2}\sum_i K_\phi(1+\cos 3\phi_i) + \sum_{i,j} V_{nb}(r_{ij}), \quad (27)$$

where $\mathbf{s} = \{b_i, \theta_i, \phi_i, \alpha_i^\sigma, \delta_i, \gamma_i^\sigma\}$ is the vector representing the internal coordinates of the atoms for a given alkane molecule, b_i are the CC bond lengths, θ_i are the CCC bond angles, ϕ_i are the CC torsional angles, α_i^σ are the

Force fields

- AMBER
- CHARMM
- GROMOS
- OPLS(-AA)

Standard force fields typically provide parameters for (in approximately descending order of quality):

- Proteins
- DNA/RNA
- Fatty acids, membranes
- Carbohydrates
- Drug-like small molecules
- Modifications of amino acids etc.

Force fields: If there are no standard parameters . . .

"Educated guess" generators and (semi-)automated optimization of parameters

- CGenFF (cgenff.paramchem.org)
- SwissParam (www.swissparam.ch)
- Automated Topology Builder (atb.uq.edu.au)
- ANTECHAMBER & GAFF (AMBER, ambermd.org)
- ffTK (www.ks.uiuc.edu/Research/vmd/plugins/fftk)
- GAAMP (gaamp.lcrc.anl.gov)
- . . .

Force fields — Dos and Donts

- Do not mix and match parameters; do not substitute a "better" water model
- Do use the available "educated guess" generators. (If you optimize an "educated guess" further or develop parameters on your own, follow the rules of the force field you are targeting to remain consistent with it.)
- Do respect cut-offs, tapering functions etc. of your force field; they are part of the parameterization

Force fields — an opinionated summary

Do blame the force field only after you have eliminated all other sources of error!

Force fields — an opinionated summary

Do blame the force field only after you have eliminated all other sources of error!

That being said, things *can* go wrong:

- Intrinsically disordered proteins (Nature Meth. 2017, 14, 71)
- Protein association (JCTC 2014, 10, 5113)

Running simulations . . .

ACEMD, AMBER, CHARMM, DESMOND, GROMACS, GROMOS, NAMD, OpenMM, . . .

Running simulations . . .

ACEMD, AMBER, CHARMM, DESMOND, GROMACS, GROMOS, NAMD, OpenMM, . . .

Programs have become more "user-friendly", and come with extensive tutorials / reasonable default settings.

- Stick with the program for which experience is available in your group!
- Avoid hunting for the "fastest" MD engine

Do not underestimate cost of MD based studies

- CPU/GPU resources for running simulations
- Disk space for storing / analyzing data
- Analysis may also be very costly!

"In working at the interface of chemistry and biology with simulation techniques, it is essential to realize that of the many exciting systems that are being studied experimentally, only relatively few pose questions for which molecular dynamics simulations can provide useful insights at their present stage of development." [MK2014][‡]

[‡] M. Karplus, Nobel lecture

Running simulations — the more you know . . .

Do study the underlying physics/algorithms $\Leftrightarrow$ *The good, the bad and the user in soft matter simulations* BBA 1858 (2016), 2529-38. *Real Cost of Speed: The Effect of a Time-Saving Multiple-Time-Stepping Algorithm on the Accuracy of Molecular Dynamics Simulations* JCTC 13 (2017), 2367-72

Running simulations — the more you know . . .

Do study the underlying physics/algorithms $\Leftrightarrow$ *The good, the bad and the user in soft matter simulations* BBA 1858 (2016), 2529-38. *Real Cost of Speed: The Effect of a Time-Saving Multiple-Time-Stepping Algorithm on the Accuracy of Molecular Dynamics Simulations* JCTC 13 (2017), 2367-72

". . . I want to caution the audience (as I always do with my students) that simulations have limitations, just as do experiments. In particular, when you appear to have discovered something new and exciting, you should be doubly careful to make certain that there is no mistake in what you have done." [MK2014]

Setting up the system

Errors/omissions during system set-up make your simulation questionable if not wrong, regardless of any computational effort!

- Any MD simulation requires reasonable starting coordinates for all atoms
  - X-ray, NMR
  - Cryo-electron microscopy
  - Integrative/hybrid (I/H) methods
  - Homology modeling
  - Assembling a larger structure from "bits and pieces"
  - ...
- What to include in the simulation?
- Reflect experimental conditions in your simulation system

Learn as much as possible about your system!

Setting up the system

Even when starting from a "traditional", experimental pdb file, you have to watch out for:

- Missing coordinates
  - Missing backbone coordinates / gaps
    $\Rightarrow$ Loop modeling
  - Missing side chain coordinates
  - Missing hydrogens
  - Quality of ligand coordinates may be doubtful

Setting up the system

Even when starting from a "traditional", experimental pdb file, you have to watch out for:

- Missing coordinates
  - Missing backbone coordinates / gaps
    $\Rightarrow$ Loop modeling
  - Missing side chain coordinates
  - Missing hydrogens
  - Quality of ligand coordinates may be doubtful
  - Ambiguities, e.g. side chain 'flips'
    $\Rightarrow$ Run WHAT_CHECK, MolProbity, NQ-Flipper etc.
- Protonation/tautomeric state(s) (protein *and* ligand!)

Setting up the system

Even when starting from a "traditional", experimental pdb file, you have to watch out for:

- Missing coordinates
  - Missing backbone coordinates / gaps
    $\Rightarrow$ Loop modeling
  - Missing side chain coordinates
  - Missing hydrogens
  - Quality of ligand coordinates may be doubtful
  - Ambiguities, e.g. side chain 'flips'
    $\Rightarrow$ Run WHAT_CHECK, MolProbity, NQ-Flipper etc.
- Protonation/tautomeric state(s) (protein *and* ligand!)
- Phosphorylation, glycolysation, ...
- "Artifacts" of experimental structures (e.g., crystal contacts)

Setting up the system

Protonation / tautomeric state

- Proteins: PROPKA (& PDB2PQR)
- Organic molecules: Various (empirical) tools, e.g., ChemAxon
- Protein–ligand complexes, e.g., Protoss/Proteins*Plus*
- **Challenge:** When assigning protonation states and choosing tautomers, your choice for one site affects (in principle) all others.

Setting up the system

Protonation / tautomeric state

- Proteins: PROPKA (& PDB2PQR)
- Organic molecules: Various (empirical) tools, e.g., ChemAxon
- Protein–ligand complexes, e.g., Protoss/Proteins*Plus*
- **Challenge:** When assigning protonation states and choosing tautomers, your choice for one site affects (in principle) all others.

Some opinionated advice

- Be cognizant of the problem/difficulties!
- Be pragmatic, e.g., focus on active site
- Document you decisions and, ideally, have someone else check
- If necessary ("strange" results), rethink your choices!
- Utilize tools for system preparation!

Illustrating some points just made (PDB: 2OJ9) . . .

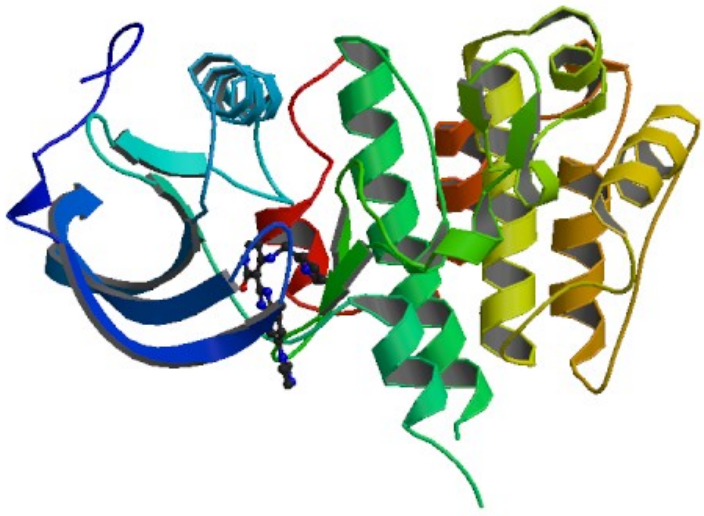
CHARMM-GUI

Effective Simulation Input Generator and More

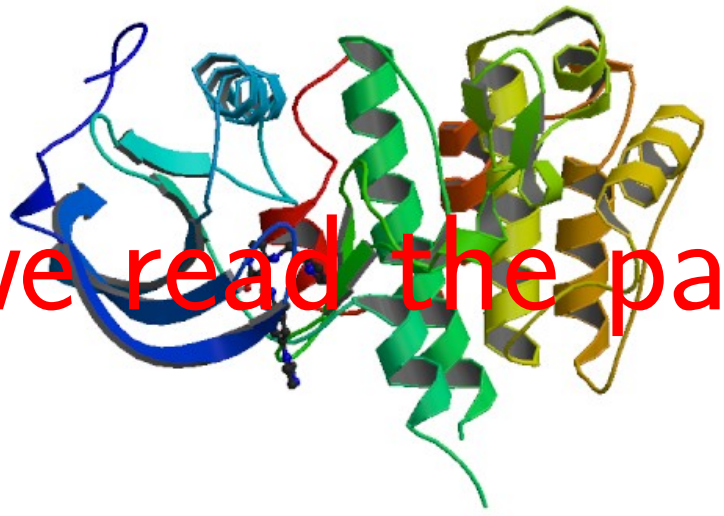
A versatile tool to set up simulations using the CHARMM force field family for various simulation programs within a few minutes

- \+ Fills gaps in structure
- \+ Automatic generation of missing parameters
- \+ Phosphorylation, glycolysation, . . .
- \+ Globular proteins, membrane proteins etc.
- – You have to decide protonation states, choose tautomers and clean up ring flips

So, how did we do (PDB: 2OJ9)?

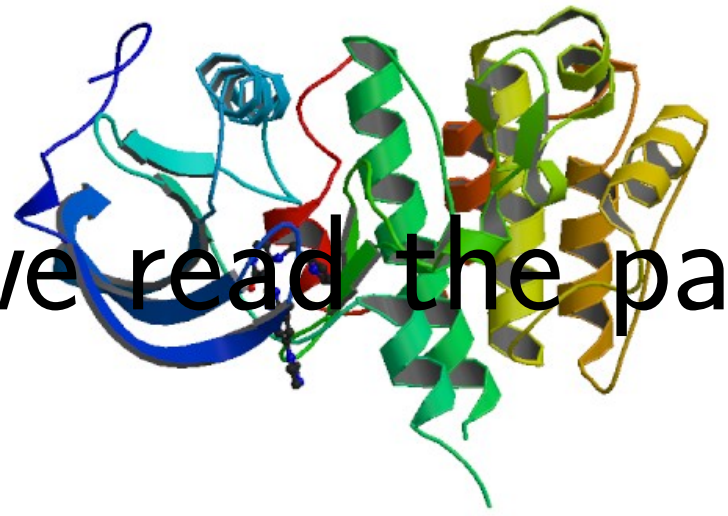


So, how did we do (PDB: 2OJ9)?

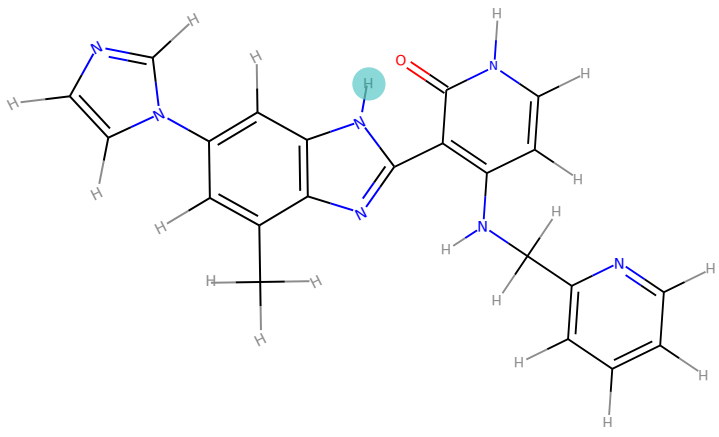


Did we read the paper??

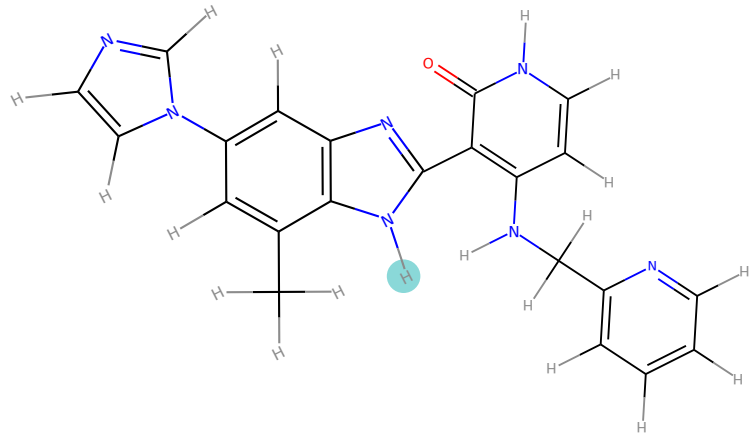
So, how did we do (PDB: 2OJ9)?



Did we read the paper??



Assumed bound form



But: structure found in PDB

Closing thoughts

- MD simulations of biomolecular systems are becoming a routine tool and are not for specialists only

Closing thoughts

- MD simulations of biomolecular systems are becoming a routine tool and are not for specialists only
- Protein Dynamics: Moore's Law in Molecular Biology (Current Biology 2011, 21, R68)

Closing thoughts

- MD simulations of biomolecular systems are becoming a routine tool and are not for specialists only
- Protein Dynamics: Moore's Law in Molecular Biology (Current Biology 2011, 21, R68)

Carrying out simulations has become "relatively easy"; thus, we can and should concentrate on carrying out meaningful simulations!

Closing thoughts

- MD simulations of biomolecular systems are becoming a routine tool and are not for specialists only
- Protein Dynamics: Moore's Law in Molecular Biology (Current Biology 2011, 21, R68)

Carrying out simulations has become "relatively easy"; thus, we can and should concentrate on carrying out meaningful simulations!

Thank you for your attention!