MD Simulations – day 1

Toni Giorgino
National Research Council of Italy
toni.giorgino@cnr.it
www.giorginolab.it

GitHub
@giorginolab

for Prof. Fuxreiter's course @ University of Padova
May 2024

https://github.com/giorginolab/MD-Tutorial-Data

This class

- Molecular dynamics is a powerful tool for studying molecular systems
- OpenMM is a software library that allows for efficient and customizable MD simulations
- It's exemplary of a modern well-maintained open-source library:
  - CI infrastructure, developed on GitHub
  - C++ w/ Python bindings
- We'll use the latter, testing *live* on Google Colab.

Molecular Dynamics

What is MD?

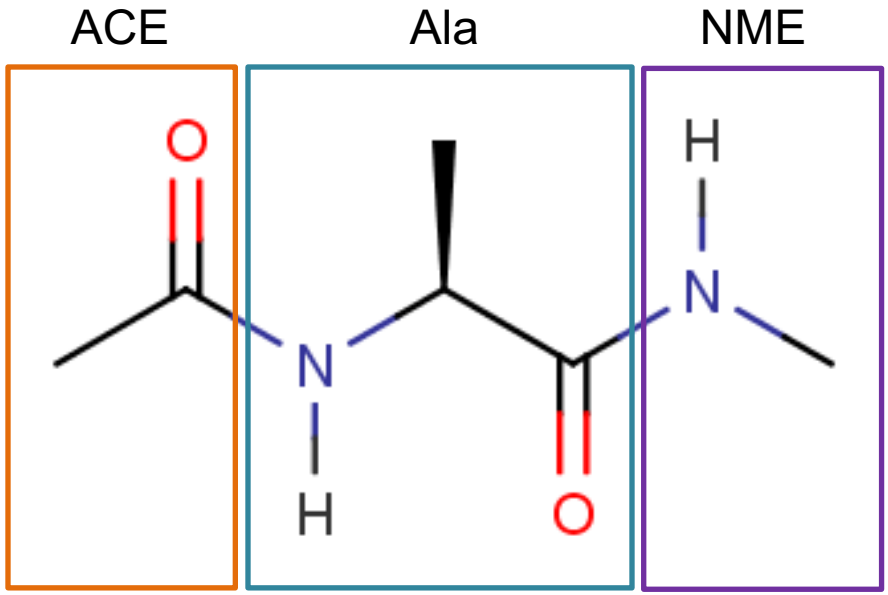
- Attempt the most detailed description of a system which is
  1. atomistic
  2. classical
- Model the internal *forces*…
- …in order to *integrate* the motion
- Hope in convergent *sampling*

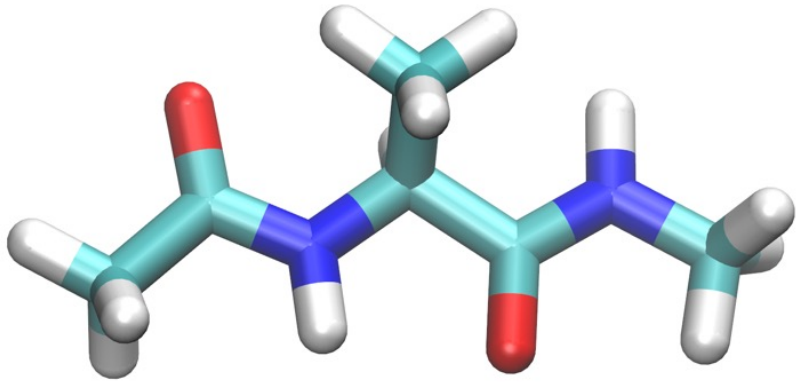
$$\vec{F}_i(\mathbf{x}) = m_i \ddot{\mathbf{x}}_i$$

Assumptions

- In this tutorial we shall deal with **unbiased** sampling approaches with **explicit** solvent, i.e.
  - no added forces except the "physical" ones in your system;
  - all of the system (including water molecules) have atomic resolution.
- Also, current classical MD does not address, by design, the following:
  - Chemical reactions, e.g. catalysis, phosphorylation, ubiquitination etc.
  - Protonation changes
- Finally, small molecules pose distinct challenges and need a separate, expensive **parameterization** step.

Alanine “dipeptide”

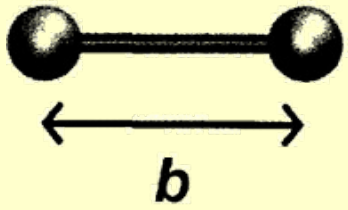


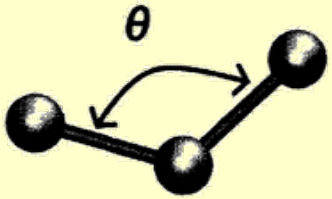


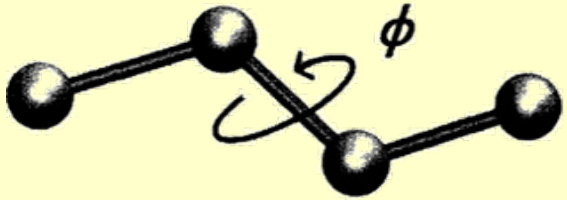
Bonded energy terms

+ Electrostatics

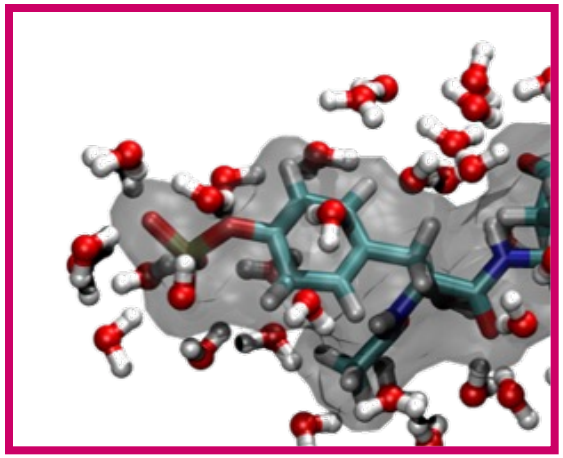
+ VdW



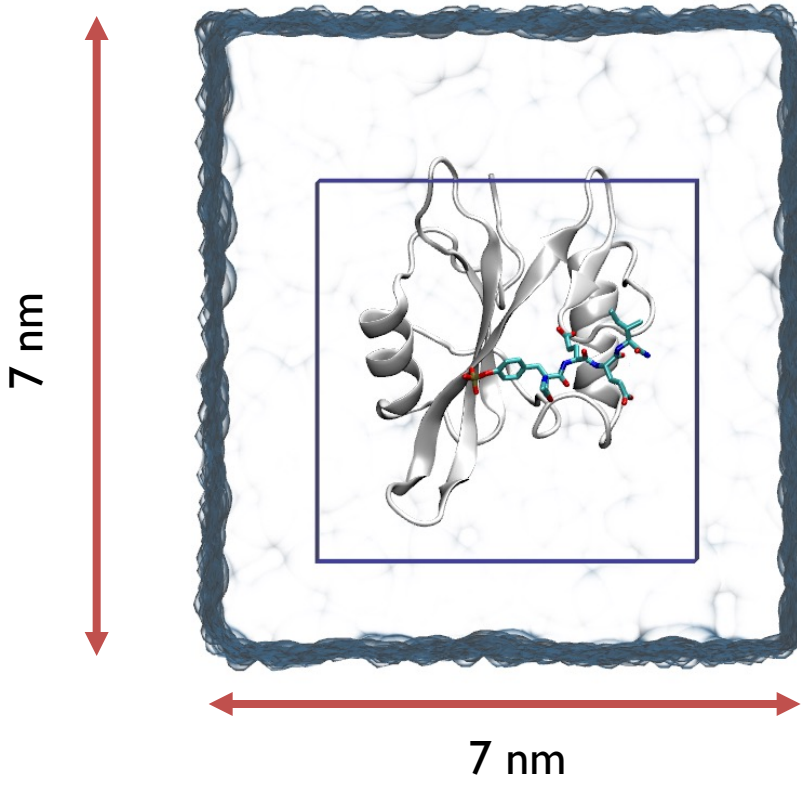


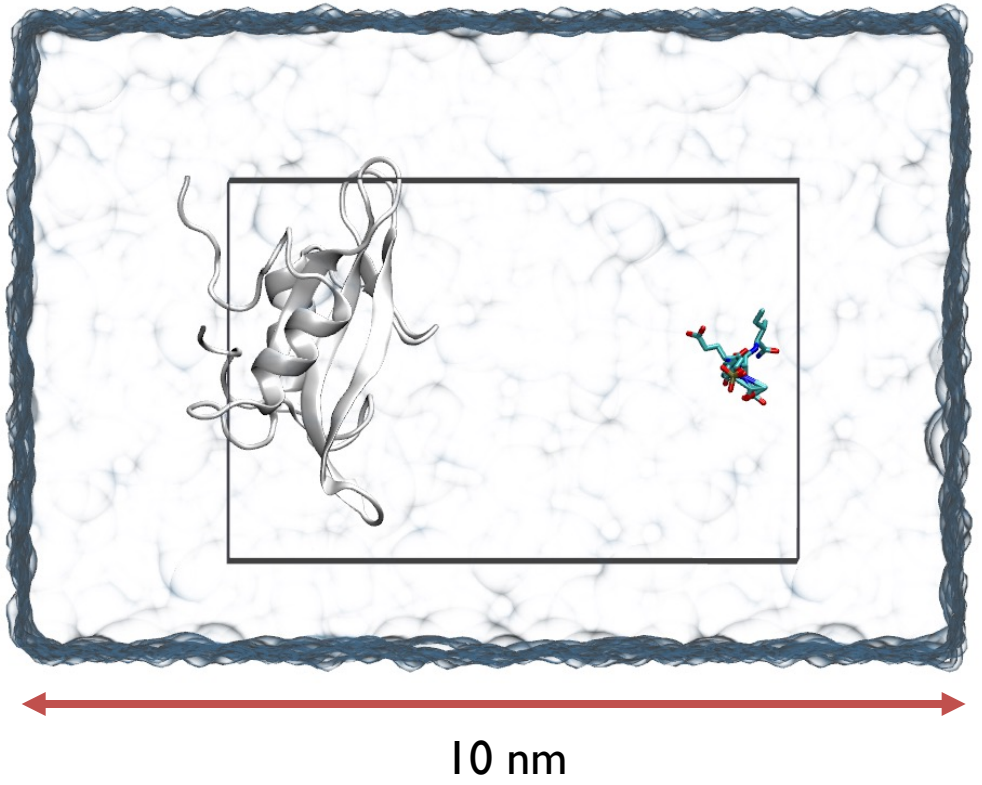


The forcefield is a database of interatomic parameters



- **Explicit solvation**
- → O(10^5) atoms
- **Unbiased dynamics**
- Update every 10^{-15} s (1 fs)





***Event* ≡ Binding / Unbinding / Folding / Unfolding / …**

* $1/t_{on}$ = association rate of SH2-pYEEI × [pYEEI]

Large gain

Ability to "play" biomolecular processes at *all-atom* resolution *in silico*

Molecular bases of folding, binding, selectivity, gating…

Large cost

E.g.*: t_{on} ~ 30 μs →

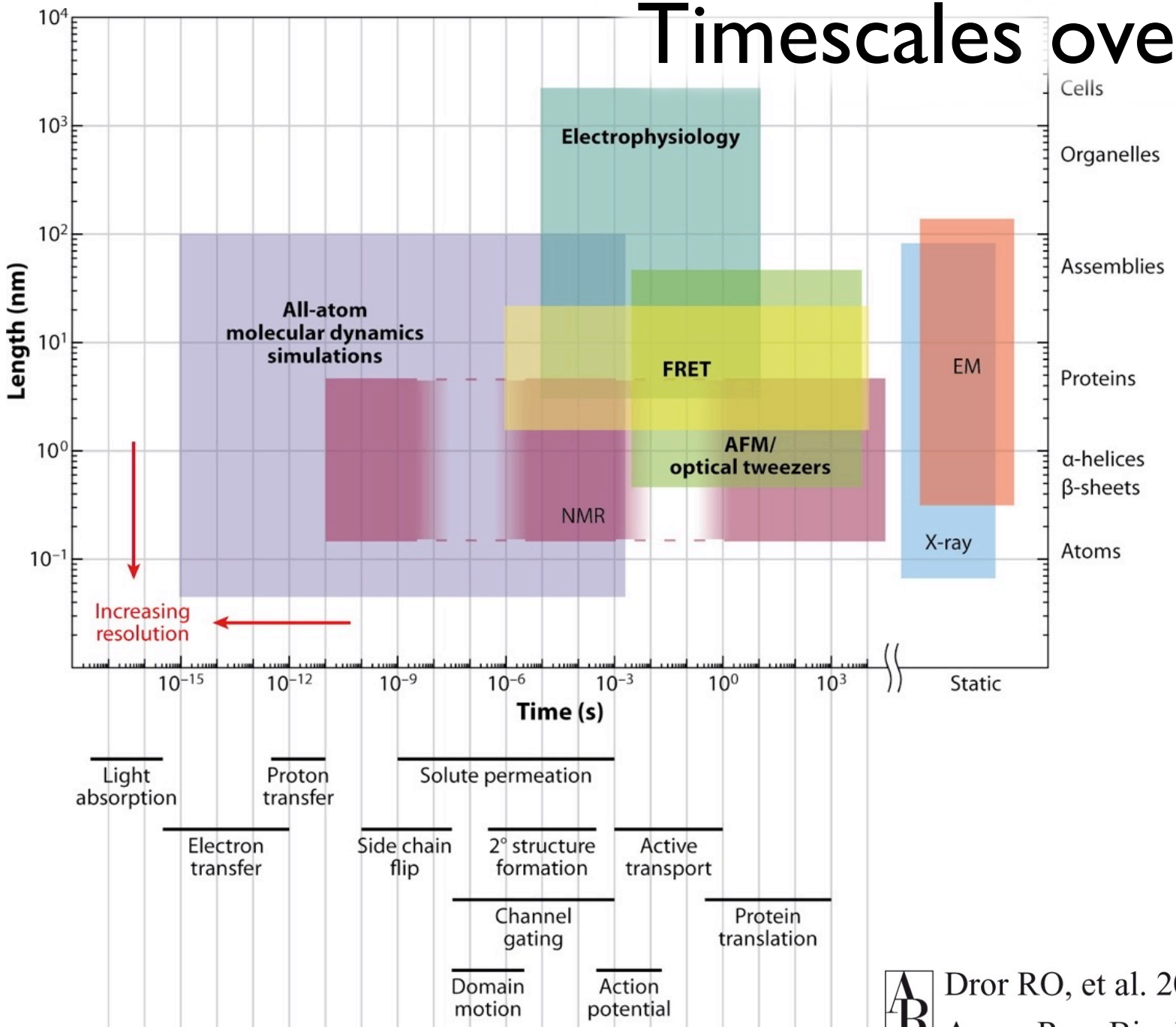
→ **10^{10}** integration timesteps →

→ 15 years single-CPU compute time

MD is entirely about timescales

- Your ability to obtain quantitative results is severely limited by the sampling ability you have. You will only be able to reach phenomena occurring on the sampled timescales, or shorter.
  - Sidechain rearrangements, diffusion-limited processes: usually possible *
  - Local flexibility: usually possible *
  - Membrane environments: ok-ish
  - Binding: hard but not impossible
  - Folding: very hard but not impossible
    - [*] Unless there are significant barriers.

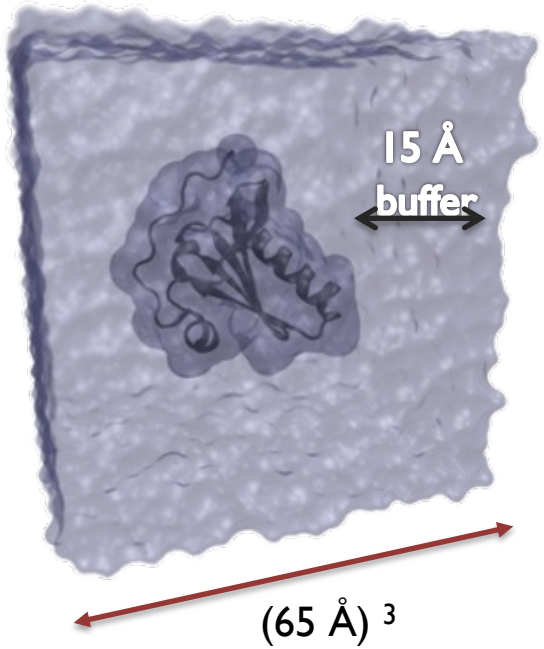
Timescales overview



Dror RO, et al. 2012.
Annu. Rev. Biophys. 41:429–52

Patience and other limits

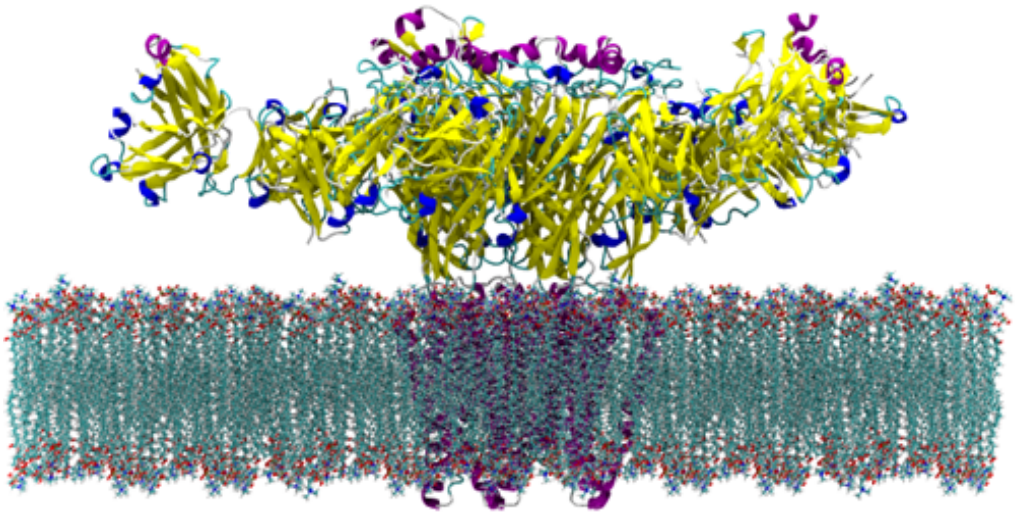
- The following factors affect the running speed (usually expressed in ns per simulation day, ns/day)
  - System size. Reasonable is 100 AA ~ 30,000 atoms.
  - Computer speed. Forget laptops.
  - Definitely use GPUs.
  - Software.



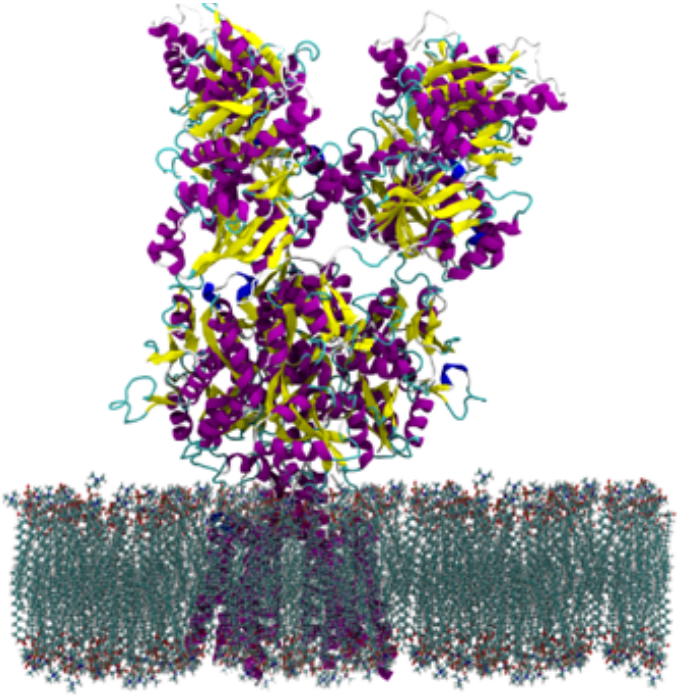
6H1F/A

”Reasonable” system sizes:

Compact
< 500 residues
< (150 A)^3
< 1 M atoms



(a) 3RHW



(b) 4U1W

Roughly in order of trouble

- Biological assembly / asymmetric unit
- Disordered regions
- Transmembrane?
- Chains
- Non standard residues
- Ligands
  - of interest
  - of no interest

- Metal ions
- Missing loops
- Termini
- Missing atoms
- Water molecules

Forcefields
and software

They run in "families"

- They are "databases" of atomic parameters
- We deal with non-polarizable all-atom ones: AMBER and CHARMM
  - There are many others, these are s.o. art
- Most notable difference:
  molecular types supported (lipids, drugs, …)
- They differ in (software) build procedures
- However OpenMM unifies them

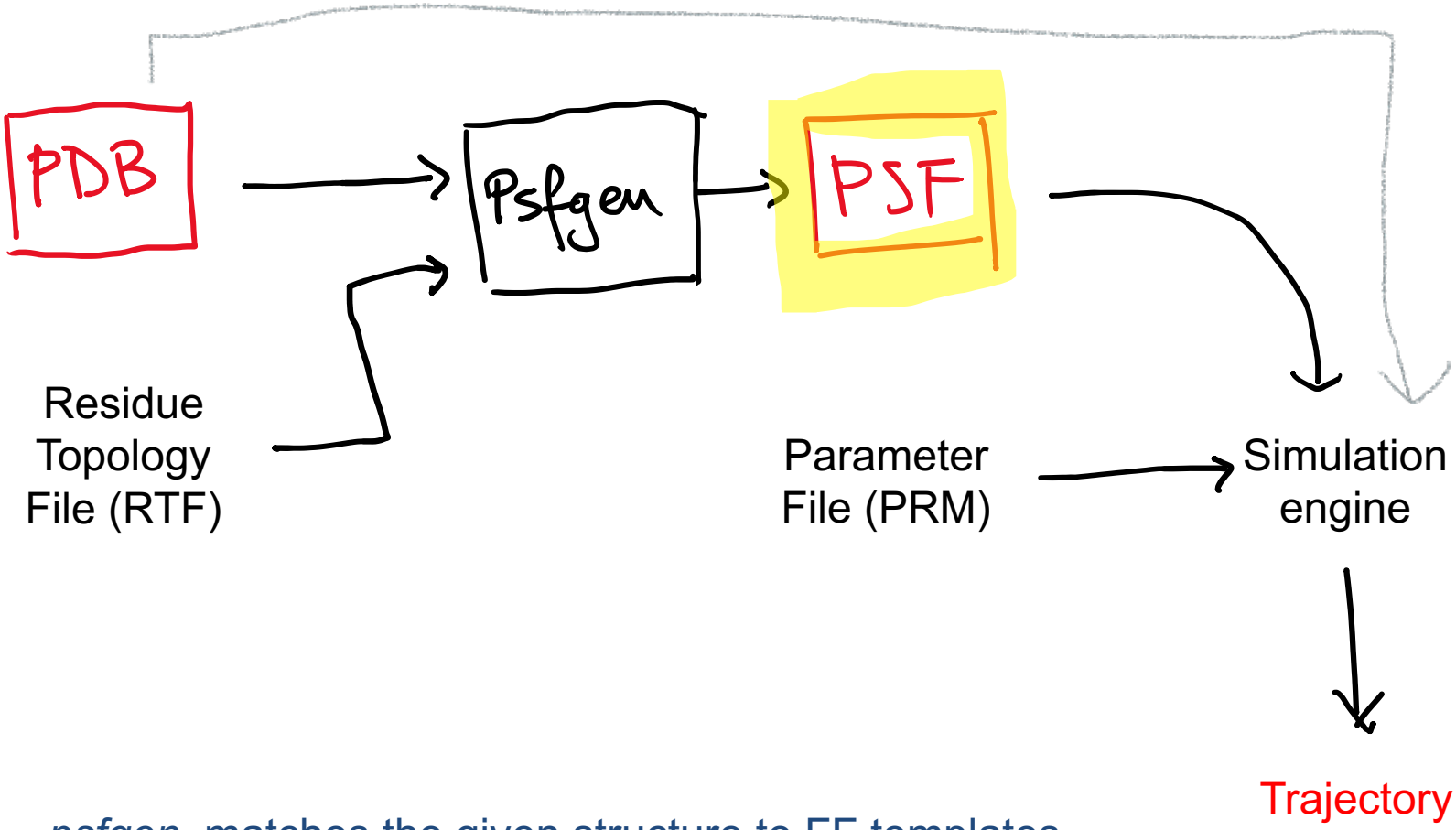
The CHARMM family

- Originally coupled with the CHARMM software (MacKerell, Karplus), but independent
- Variants of note: C36M
  - toppar_c36_jul22.tgz
- Based on RTF (templates) and PRM (parameters) files
- Build: **psfgen** -> xxx.psf

- https://www.academiccharmm.org/
- https://mackerell.umaryland.edu/charmm_ff.shtml

CHARMM
Chemistry at HARvard Macromolecular Mechanics

CHARMM system layout

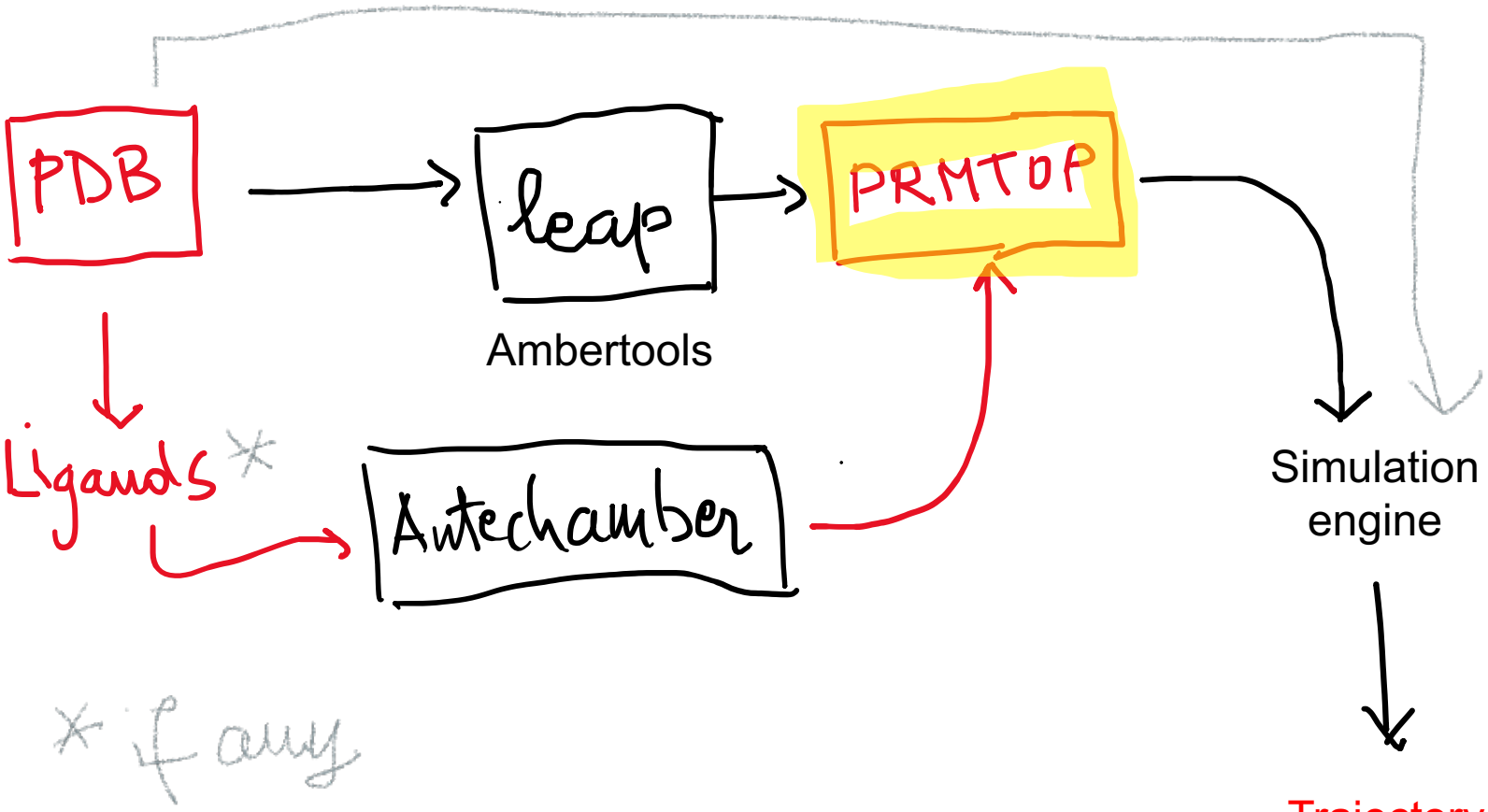


psfgen matches the given structure to FF templates provided in RTF files, completing missing atoms. The actual parameter values are read at a later stage.

The AMBER family

- Originally coupled with the Amber software (Case, Merz, Kollmann, …), but independent
- Variants of note: ff19SB
  - https://ambermd.org/AmberTools.php
- Based on **tleap** + its database
- Build: **tleap** --> xxx.prmtop

Amber system layout



tleap matches the given structure to FF templates, completing missing atoms, *merging parameters*.

FF choice

Largely equivalent, i.e. subtle differences only appear at late stages.
FF choice is mostly due to the species in the system that one intends to model.

	CHARMM	AMBER
Proteins, peptides	++	++
Water, ions	++	++
Lipids (membranes)	++	+
Small molecules	+ (CGenFF)	++ (GAFF2)
Post-translational modif. *	+/-	+/-
Non-standard charge states	+/-	+/-
DNA	-	-
RNA	--	--
Non-standard AAs *	--	--

Somewhat subjective!

* Check individually

An example for Amber

Molecule/Ion Type	Force Field
protein	ff19SB
DNA	OL21
RNA	OL3
carbohydrates	GLYCAM_06j
lipids	lipids21
organic molecules (usually ligands)	gaff2
ions	•should be matched to water model; see force fields for ions for further discussion
water model	•should be matched to atomic ions; common water models include tip3p, spc/e, tip4pew, and OPC

System building

The build procedure

- Know your system
- Build
  - Cleanup
  - Assign forcefield
  - Solvate
  - Ionize
- Minimize
- Equilibrate
- Run

The build procedure

- It used to be somewhat convolved
- Generally
  - take PDB coordinates
  - filter out unwanted species
  - solvate
  - ionize
- Automation was (it still is) challenging
- OpenMM unified the build process (but it is still possible to use the old tools)

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

High-Throughput Automated Preparation and Simulation of Membrane Proteins with HTMD

Stefan Doerr,[†,∥] Toni Giorgino,[‡,∥] Gerard Martínez-Rosell,[†] João M. Damas,[¶] and Gianni De Fabritiis*[,†,§]

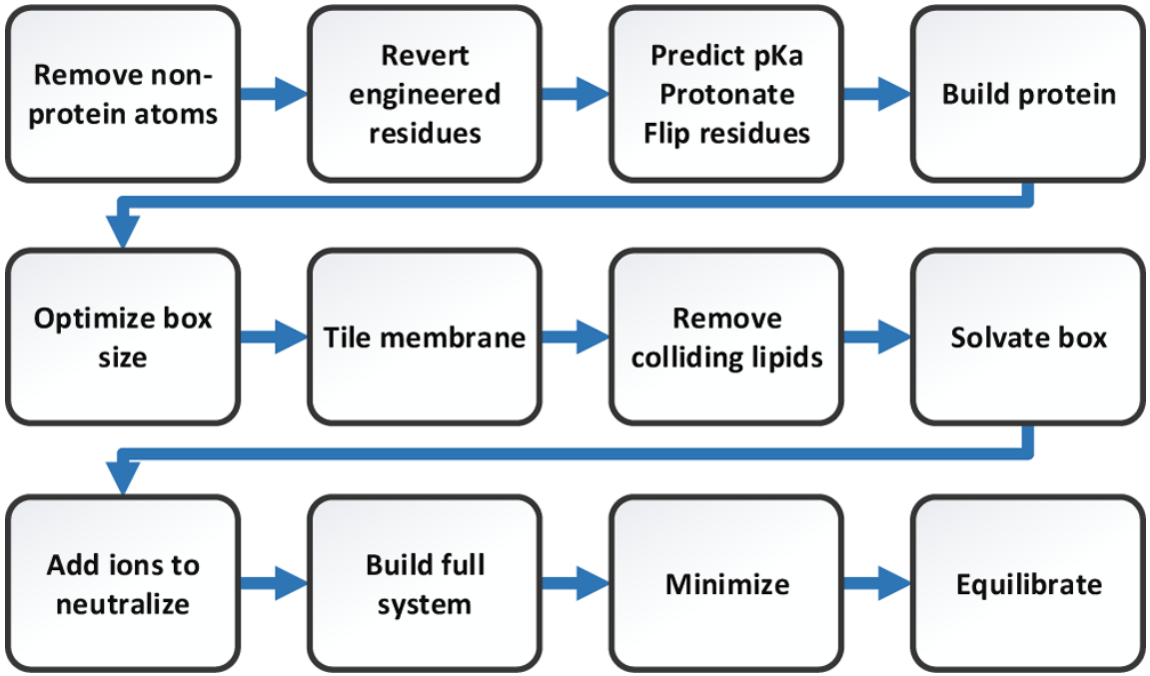


Figure 2. OPM building and simulation protocol workflow starting from a PDB file and ending with an equilibrated system.

Small molecules
in the Amber system

Ligands

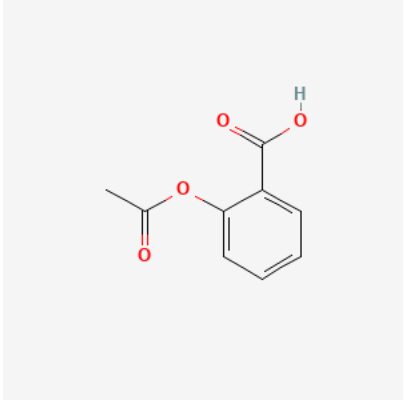
- Anything non-peptide is problematic
- Some common molecules are pre-computed
- Generic small molecules need *parameterization*
- AmberTools has *antechamber* + *GAFF*
  - Semi-automated
  - Do mind stereochemistry, tautomers, protonation!
  - Partial charges are assigned by RESP
  - Other force terms are pattern-matched
- CHARMM has CGenFF, web only

Ligands

CC(=O)OC1=CC=CC=C1C(=O)O

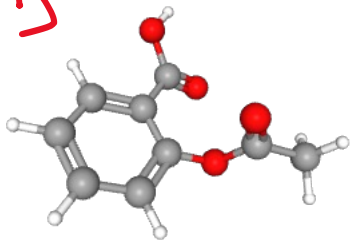
Smiles

(databases for virtual screening)



planar

(editing)



3D conformers

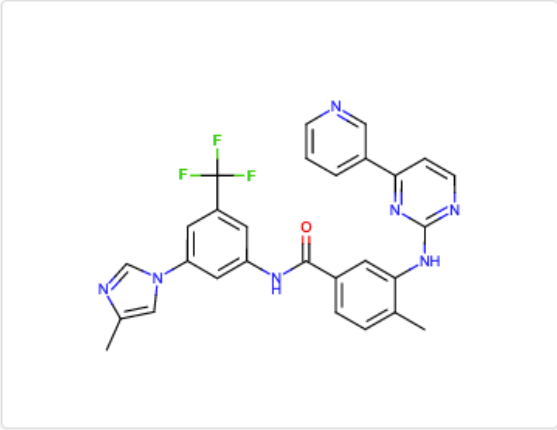
generated or crystallized

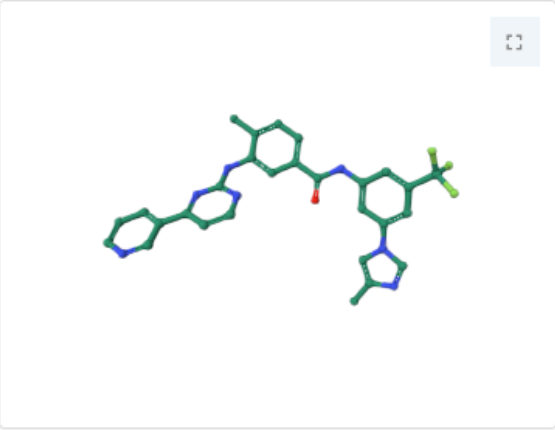
- .mol2
- .sdf

further process

(dock, parameterize)

RCSB PDB "Ligand expo"





Toggle Hydrogen | Toggle Labels

Display Files | Download Files | Data API

NIL

Nilotinib

Find entries where: NIL

☑ is present as a standalone ligand in 3 entries

search

Find related ligands:

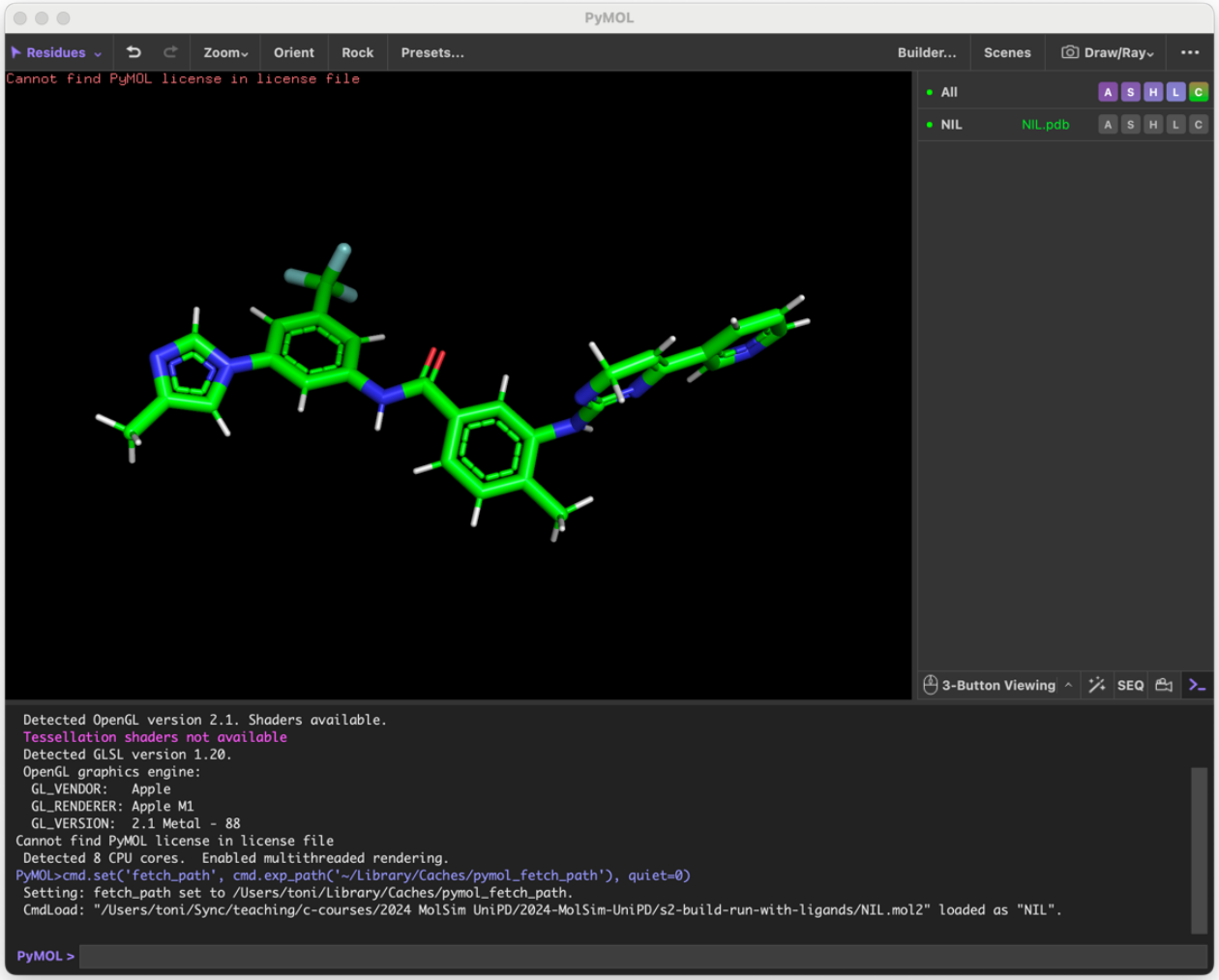
Similar Ligands (Stereospecific)
Similar Ligands (including Stereoisomers)
Similar Ligands (Quick Screen)
Similar Ligands (Substructure Stereospecific)
Similar Ligands (Substructure including Stereoisomers)

Chemical Component Summary

Name	Nilotinib
Synonyms	4-methyl-N-[3-(4-methyl-1H-imidazol-1-yl)-5-(trifluoromethyl)phenyl]-3-[(4-pyridin-3-ylpyrimidin-2-yl)amino]benzamide
Identifiers	4-methyl-N-[3-(4-methylimidazol-1-yl)-5-(trifluoromethyl)phenyl]-3-[(4-pyridin-3-ylpyrimidin-2-yl)amino]benzamide
Formula	C_{28} H_{22} F_3 N_7 O
Molecular Weight	529.516
Type	NON-POLYMER

Chemical Details

Formal Charge	0
Atom Count	61
Chiral Atom Count	0
Bond Count	65
Aromatic Bond Count	31


PyMOL
Residues
Zoom
Orient
Rock
Presets...
Builder...
Scenes
Draw/Ray
Cannot find PyMOL license in license file
All
NIL
NIL.pdb
3-Button Viewing
SEQ
Detected OpenGL version 2.1. Shaders available.
Tessellation shaders not available
Detected GLSL version 1.20.
OpenGL graphics engine:
GL_VENDOR: Apple
GL_RENDERER: Apple M1
GL_VERSION: 2.1 Metal - 88
Cannot find PyMOL license in license file
Detected 8 CPU cores. Enabled multithreaded rendering.
PyMOL>cmd.set('fetch_path', cmd.exp_path('~/Library/Caches/pymol_fetch_path'), quiet=0)
Setting: fetch_path set to /Users/toni/Library/Caches/pymol_fetch_path.
CmdLoad: "/Users/toni/Sync/teaching/c-courses/2024 MolSim UniPD/2024-MolSim-UniPD/s2-build-run-with-ligands/NIL.mol2" loaded as "NIL".
PyMOL >

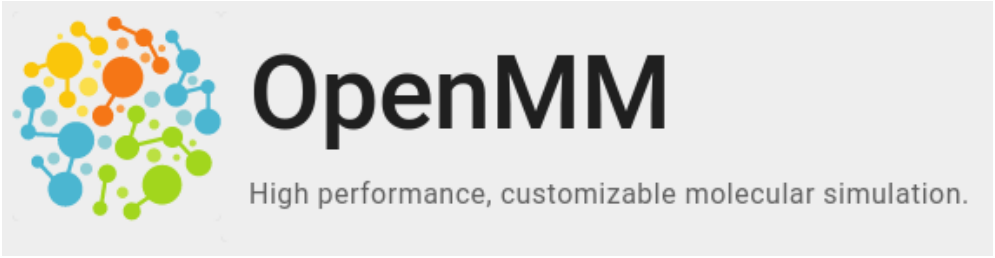
Practice, finally

Open these

- www.rcsb.org/structure/3CS9
- github.com/giorginolab/2024-MolSim-UniPD

2024-MolSim-UniPD / amber-build-run-with-ligand /

OpenMM



.org

- OpenMM is a molecular dynamics simulation toolkit that allows for high-performance simulations of biomolecules.
- Allows for simulation of a variety of molecular systems, including proteins, nucleic acids, and small molecules
- OpenMM supports a wide range of force fields and integrators and can run on CPUs and GPUs.
- Open source, written in C++ with Python and other language bindings available

Basic Workflow (object-oriented)

1. Download, complete and edit the structure:
   - **Topology** (i.e. the identity of atoms, bonds, etc)
   - **Positions** (i.e. the starting coordinates)
2. Create the **system** object.
3. Create the **integrator** object.
4. Create and add custom **forces** to system if needed.
5. Define the **simulation** object.
6. Set the initial positions and velocities.
7. Minimize.
8. Run the simulation.
9. (Analyze the results.)

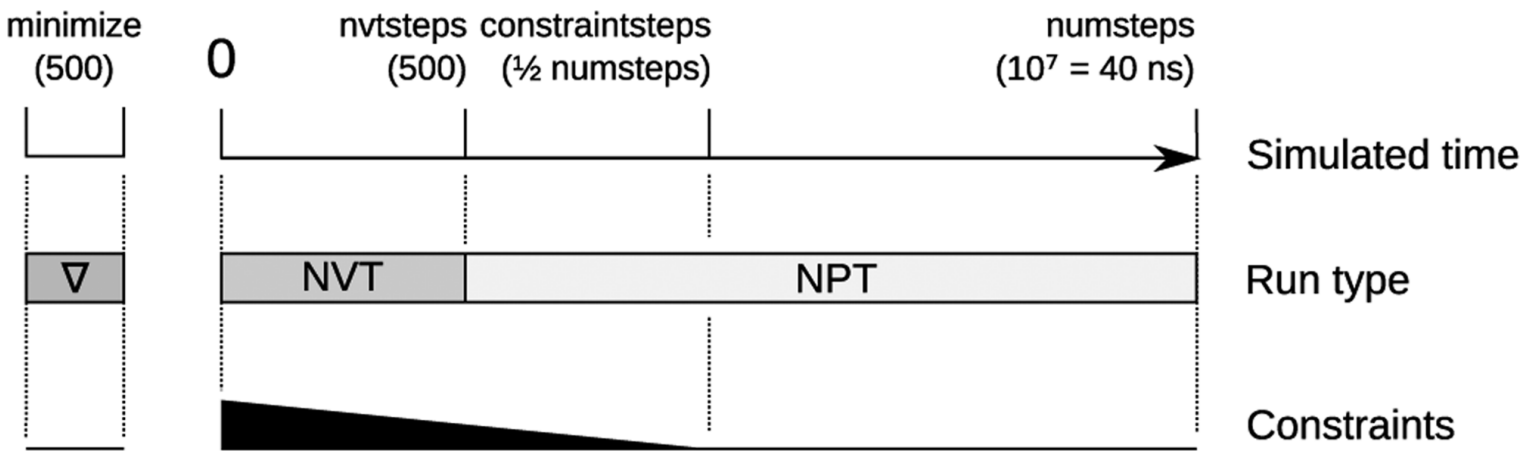
Integrators

- …are algorithms that solve the equations of motion for a system
- OpenMM includes several integrators, e.g. Langevin dynamics, Verlet integrator, and Monte Carlo barostat
- Different integrators are appropriate for different types of simulations and conditions (e.g.: NPT vs NVT)

Simulating a system

- Once a system has been defined and the force field and integrator selected, it can be simulated
- The simulation (run) involves running a series of steps, where each step involves calculating the forces on each atom, integrating the equations of motion, and updating the system's coordinates
- After the simulation, data analysis can be performed to obtain information about the system's behavior and properties

Equilibration



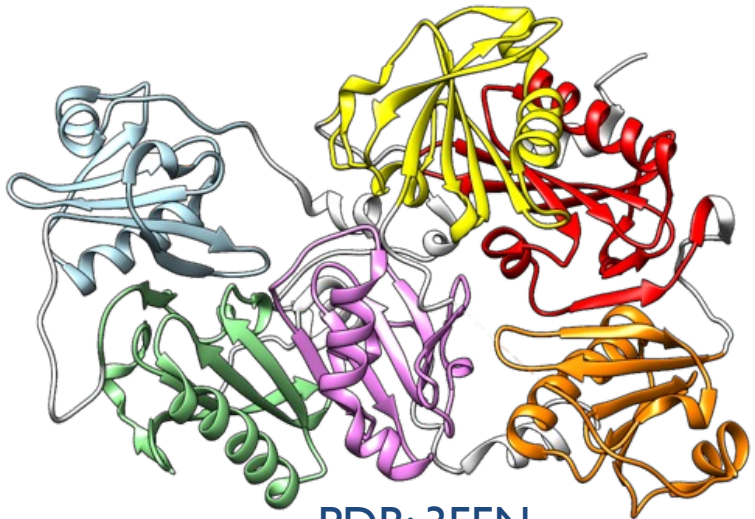
Orders of magnitude

- Integration timestep: 1-4 fs
  - How often forces are calculated, i.e. the finest time granularity. No periods faster than this.
- Logging interval: O(1 ps)
  - How often to print energies etc.
- Frame interval: O(100 ps)
  - How often to save a snapshot
- Total length: O(10-1000 ns)

Let’s pick a test system

Inactive

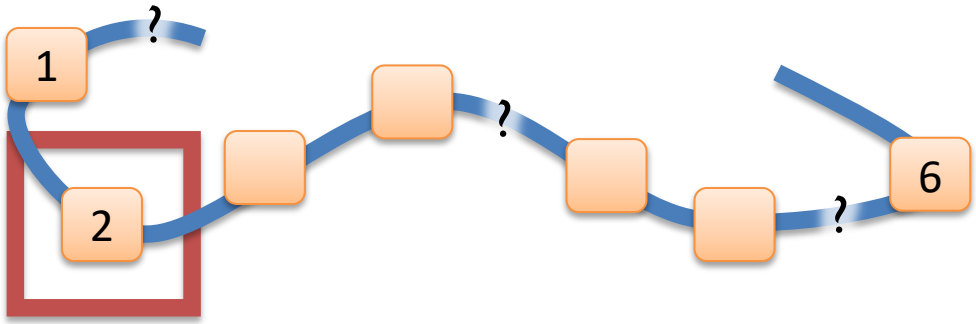
No Ca^{2+}.
Closed form.
Resolved (2008).



~ 100 Å

Active

~ mM Ca^{2+}.
Active form.
Dynamic. Elusive.



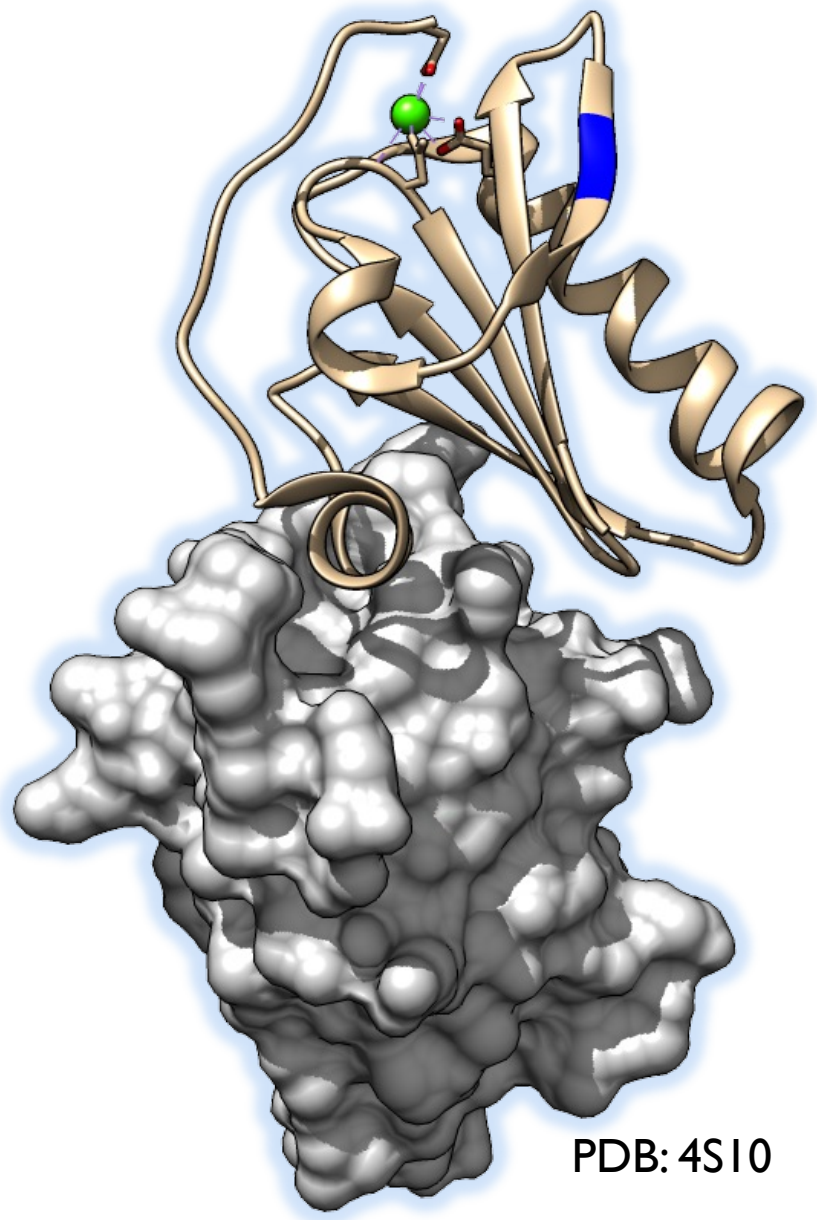
However, crystallisation of Ca^{2+}-bound **isolated domains** has been successful.

2015: Gettemans et al. – G2 Nanobodies

- Inoculate llamas with GSN **WT** G2
- Extract nanobodies (Nb11)
- Sequence them
- Obtain WT:Nb structure

Idea

- Can the WT-raised Nb11 re-stabilize D187N enough to allow crystallization?



PDB: 4S10

Van Overbeke et al., HMG 2015

6H1F: Gelsolin G2+nanobody

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome | Versions

Biological Assembly 1

3D View: Structure | 1D-3D View | Electron Density | Validation Report | Ligand Interaction

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Hetero 2-mer - A1B1

Find Similar Assemblies

Biological assembly 1 assigned by authors and generated by PISA (software)

Biological Assembly Evidence: gel filtration

Macromolecule Content

- Total Structure Weight: 28.49 kDa
- Atom Count: 1,896
- Modelled Residue Count: 229
- Deposited Residue Count: 259
- Unique protein chains: 2

Display Files | Download Files

6H1F

Structure of the nanobody-stabilized gelsolin D187N variant (second domain)

PDB DOI: 10.2210/pdb6H1F/pdb

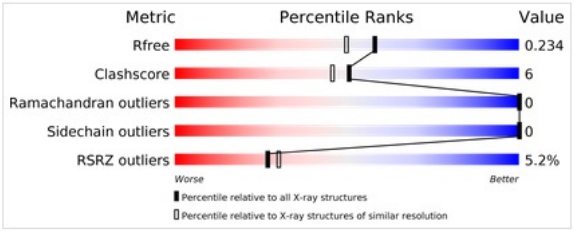
Classification: STRUCTURAL PROTEIN
Organism(s): Lama glama, Homo sapiens
Expression System: Escherichia coli
Mutation(s): Yes

Deposited: 2018-07-11 **Released:** 2019-01-23
Deposition Author(s): Hassan, A., Milani, M., Mastrangelo, E., de Rosa, M.
Funding Organization(s): Amyloidosis Foundation

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.90 Å
R-Value Free: 0.233
R-Value Work: 0.199
R-Value Observed: 0.202

wwPDB Validation | 3D Report | Full Report



This is version 1.0 of the entry. See complete history.

Literature | Download Primary Citation

Nanobody interaction unveils structure, dynamics and proteotoxicity of the Finnish-type amyloidogenic gelsolin variant.

Giorgino, T., Mattioni, D., Hassan, A., Milani, M., Mastrangelo, E., Barbiroli, A., Verhelle, A., Gettemans, J., Barzago, M.M., Diomede, L., de Rosa, M.
(2019) Biochim Biophys Acta Mol Basis Dis **1865**: 648-660

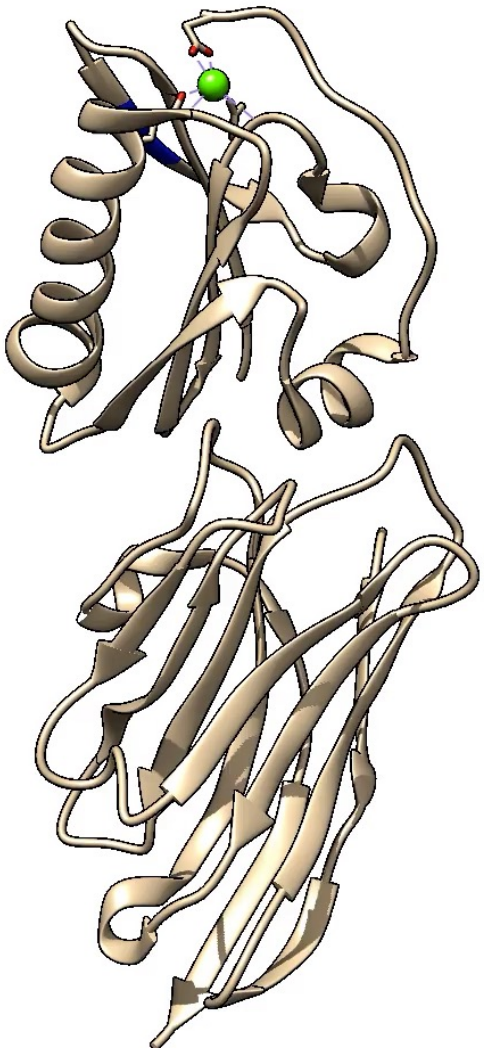
PubMed: 30625383 Search on PubMed
DOI: 10.1016/j.bbadis.2019.01.010
Primary Citation of Related Structures:

Three puzzles!

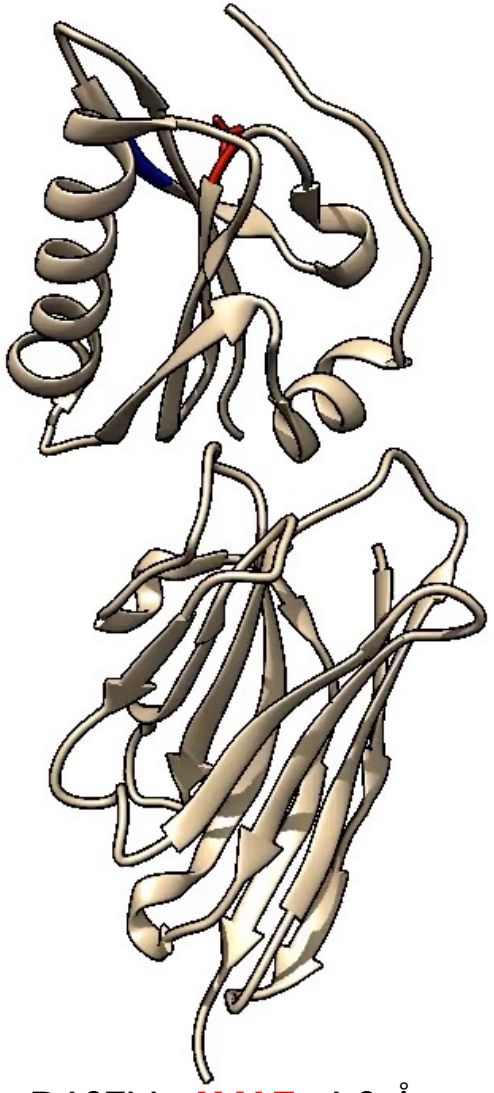
1. WT and D187N are **virtually identical***: same structure, different function
2. Nb11 binds far from the furin cleavage site…
3. …and far from the Ca^{2+} ion

* Except Ca^{2+} binding

WT:Nb11 complex compared to D187N:Nb11.



WT: 4S10, 2.6 Å



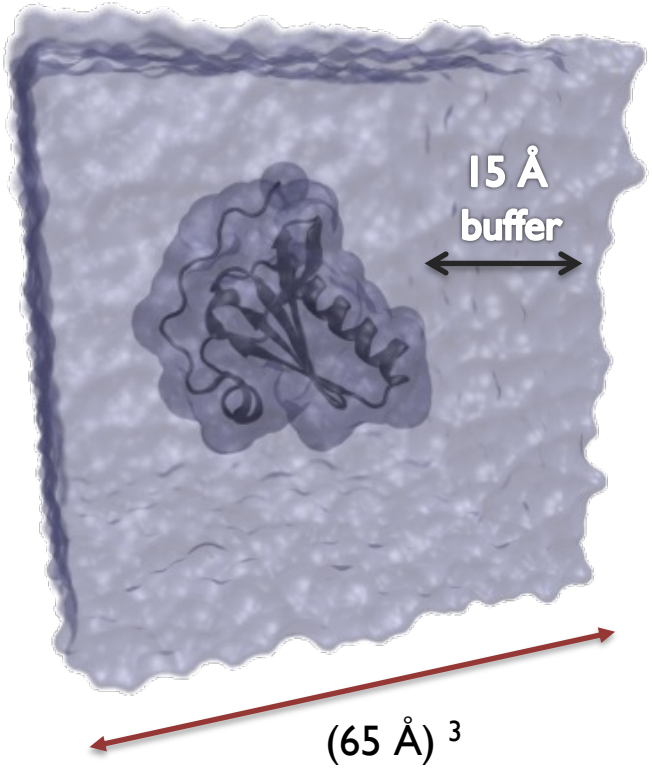
D187N: **6H1F**, 1.9 Å

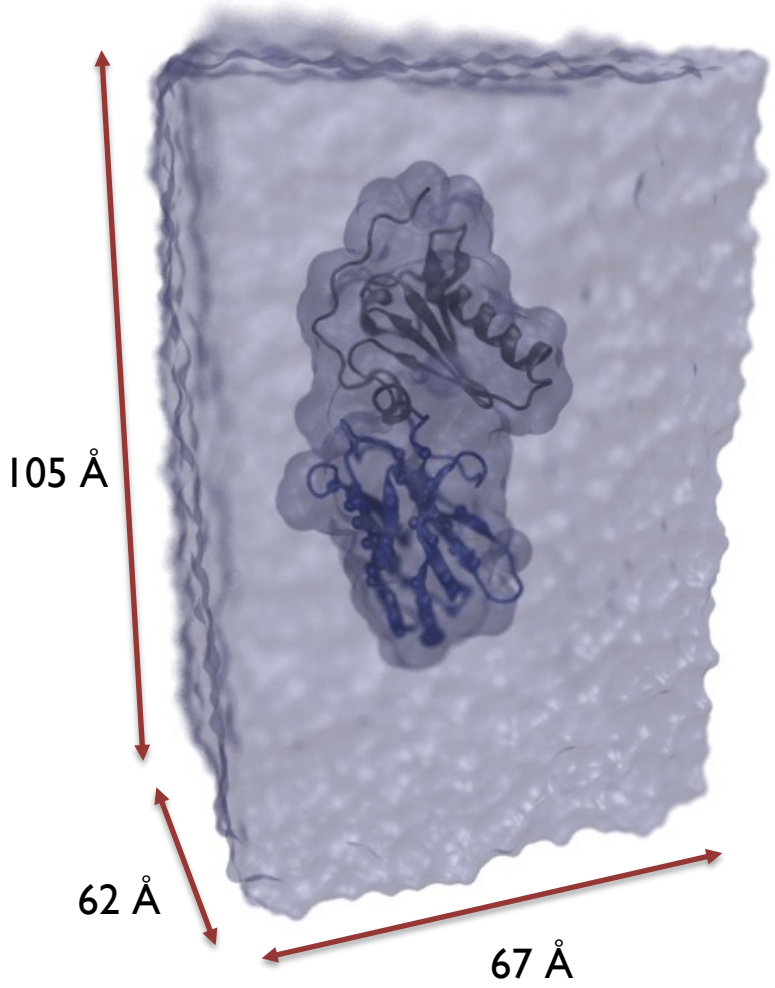
GSN ± Nb11 MD simulations

- Unbiased sampling @300 °K
- 100 mM NaCl
- Harmonic restraints:
  SS Nb1 @ 0.03 kcal/mol/Å^2

CHARMM36

~3 μs tot. ~25k/43k atoms





Open these

- www.rcsb.org/structure/6H1F
- github.com/giorginolab/2024-MolSim-UniPD

2024-MolSim-UniPD / openmm-python-only / **OpenMM_2024.ipynb**

The code contained in this repo is structure in large part as IPython notebooks. They can be run in three ways:

- On your local machine. That's probably the easiest and most didactic for interpreting the intermediate files.
- On Google Colab. Just open Colab and point it to the file's URL. Simulations run much faster if you request a GPU runtime.
- On Binder. Somewhat slow but comes with a pre-installed environment. Open in binder

Results

- The results from actual (long) runs are in
  github.com/giorginolab/MD-Tutorial-Data

- Next week will do the analysis etc., but we can already visualize and discuss.

Conclusion

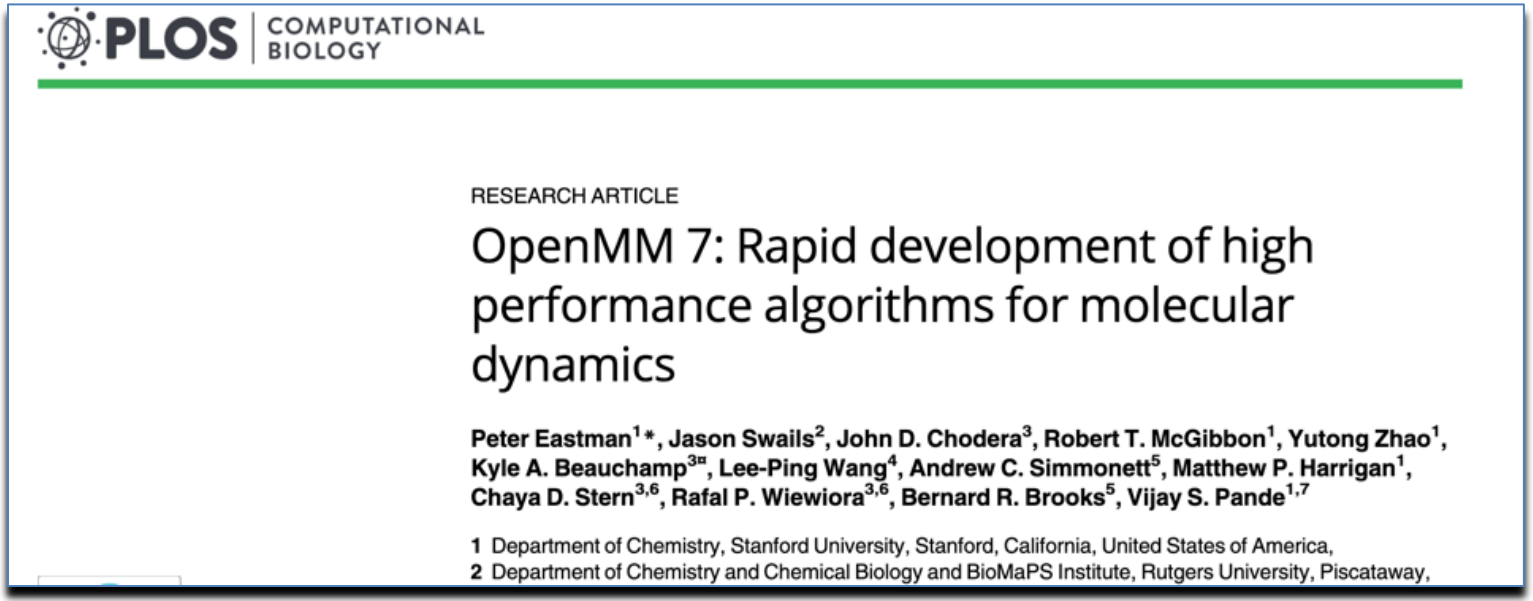
Conclusion

- OpenMM is a powerful tool for molecular dynamics simulations
- Good, if fragmented, documentation
- With its customizable force fields and integrators, it can be used to study a wide range of atomistic systems, e.g.
  - "toy" polymers
  - all-atom MD with major FFs
  - ANN potentials

Resources for learning OpenMM

- OpenMM.org website and documentation
- GitHub repository with examples and tutorials
- Community forums and mailing lists for support and discussion

- See also
  - OpenMMtools
  - https://openforcefield.org/
  - HTMD, ACEMD
  - https://github.com/openmm/pdbfixer
  - Charmm-GUI



PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

OpenMM 7: Rapid development of high performance algorithms for molecular dynamics

Peter Eastman[1]*, Jason Swails[2], John D. Chodera[3], Robert T. McGibbon[1], Yutong Zhao[1], Kyle A. Beauchamp[3¤], Lee-Ping Wang[4], Andrew C. Simmonett[5], Matthew P. Harrigan[1], Chaya D. Stern[3,6], Rafal P. Wiewiora[3,6], Bernard R. Brooks[5], Vijay S. Pande[1,7]

1 Department of Chemistry, Stanford University, Stanford, California, United States of America,
2 Department of Chemistry and Chemical Biology and BioMaPS Institute, Rutgers University, Piscataway,

End