Setting up MD Simulations of Biomolecules

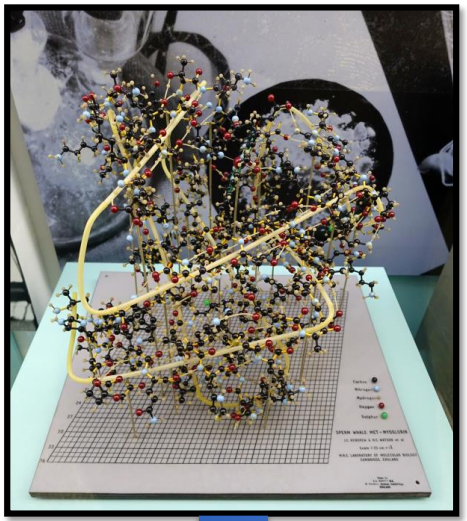
Stefan Boresch*

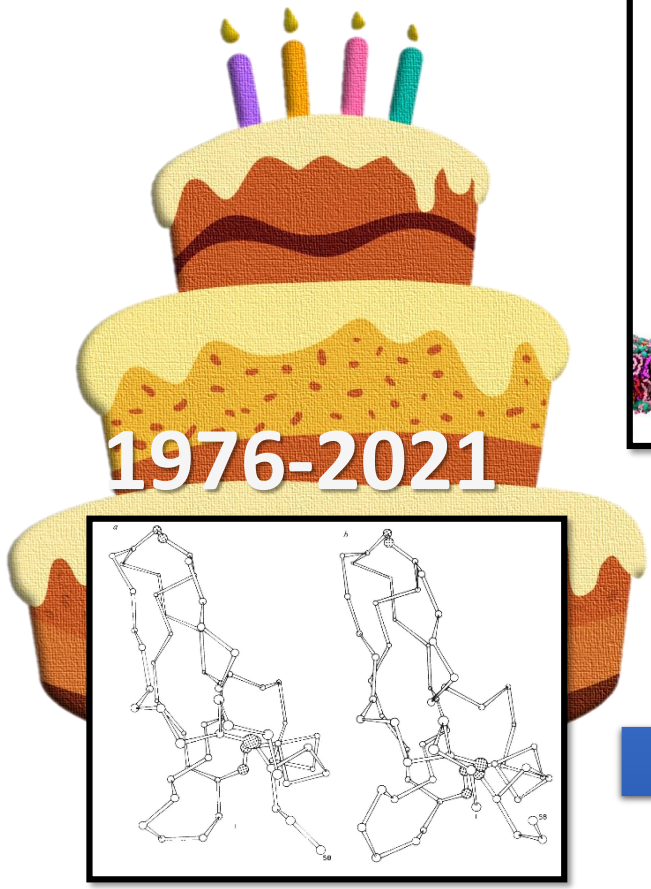
(stefan@mdy.univie.ac.at)

Virtual Europin 2021

*Univ. Vienna, Faculty of Chemistry, Austria

Happy Birthday, biomolecular MD!!!





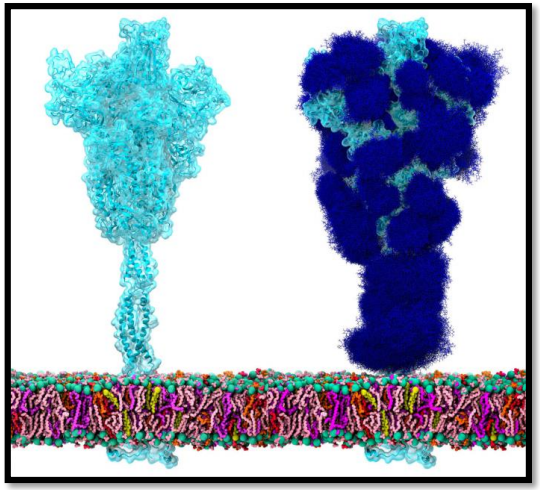
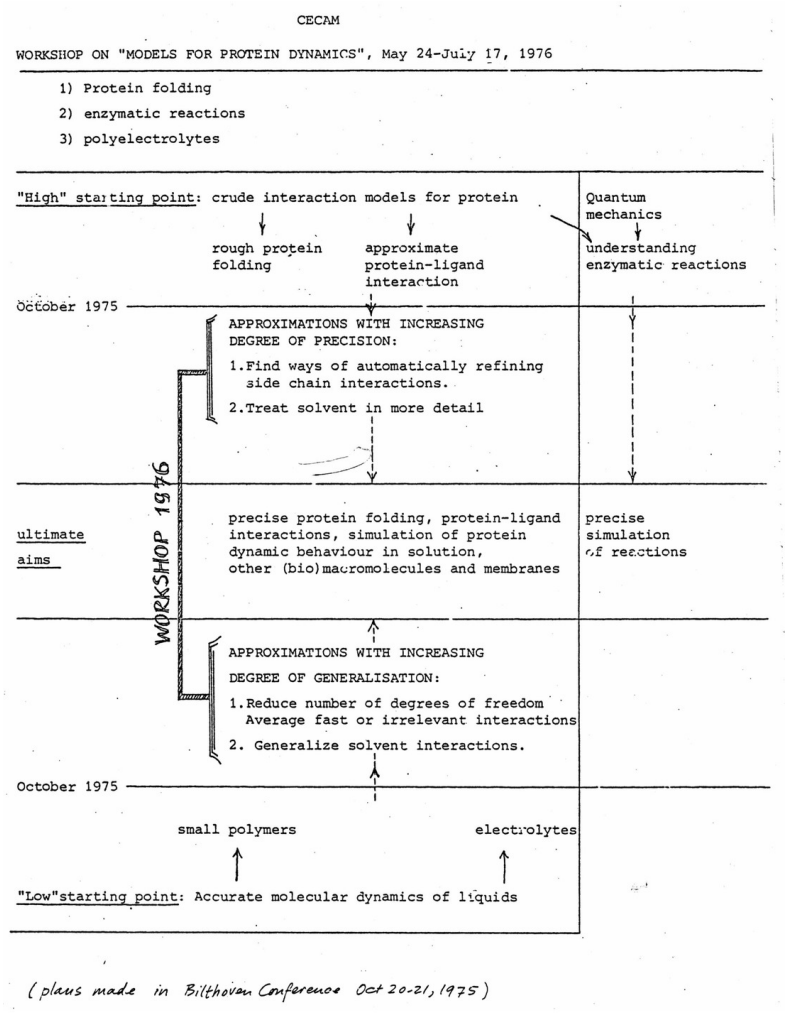


Image credits: Cartoon (clipartmax.com), 1965 myoglobin model (wikimedia, BPTI (McCammon, Gelin & Karplus, Nature 1977), spike protein (Amaro lab, via NYTimes article), collage: Clara Boresch

Models for Protein Dynamics, CECAM workshop May/June 1976



LIST OF PARTICIPANTS

C.Bennett, IBM Watson Research Center, Yorktown Heights, New York,
H.J.C.Berendsen, Lab.of Physical Chemistry, the University of Groningen, the Netherlands,
G.Careri *, Istituto di Fisica "G.Marconi", Roma,
G.Ciccotti, Istituto di Fisica "G.Marconi", Roma,
C.Chothia, Institut Pasteur, Paris,
D.Elkkoubi, Lab. d'Electrochimie, Université P. et M. Curie, Paris,
A.Englert, Chimie générale, Université Libre, Bruxelles,
D.L.Ermak, Lawrence Livermore Lab., Livermore, California,
*D.R.Ferro**, Istituto di Chimica della Macromolecule, Milano, Italia,
W.F.van Gunsteren, Lab.of Physical Chemistry, the University of Groningen, the Netherlands,
J.Hermans, Dpt. of Biochemistry, University of North Carolina, Chapel Hill, North Carolina,
*M.Karplus***, Dpt. of Chemistry, Harvard University, Cambridge, Mass.,
M.Leclerc, Chimie générale, Université Libre, Bruxelles,
*M.Levitt**, MRC Lab. of Molecular Biology, Cambridge, England,
B.Maigret, Institut de Biologie Physico-Chimique, Paris V,
J.A.Mc.Cammon, Harvard University, Cambridge, Mass.,
K.Nagano, University of Tokyo, Japan,
J.Orban, Faculté des sciences, Université Libre, Bruxelles,
S.Prémilat, Lab. de Biophysique, Nancy, France,
A.Rahman, Argonne Natl. Lab., Argonne, Illinois,
P.Rossky, Harvard University, Cambridge, Mass.,
J.P.Rijckaert, Faculté des sciences, Université Libre, Bruxelles,
P.Turq, Lab. d'Electrochimie, Université P. et M. Curie, Paris,
S.Wodak, Dept. of Chemical Biology, Université Libre, Bruxelles.

*for part of the workshop
**visitor

Note. Many reports contain some work done in a few months after the workshop ended.
In some cases this has involved cooperation with non-participants of the workshop, who are then listed as coauthors of the report.

MD in computational chemistry and biology

- Statistical Mechanics at atomic resolution, structure *and* dynamics:

  There was a sense, even at the time, of something truly historic going on, of getting these first glimpses of how an enzyme molecule for example, might undergo internal motions that allow it to function as a biological catalyst. (J.A.McCammon, Oral History (1995))

MD in computational chemistry and biology

- Statistical Mechanics at atomic resolution, structure *and* dynamics:

  There was a sense, even at the time, of something truly historic going on, of getting these first glimpses of how an enzyme molecule for example, might undergo internal motions that allow it to function as a biological catalyst. (J.A.McCammon, Oral History (1995))

  $\Rightarrow$ MD as "computational microscope" (e.g., see here and here)

MD in computational chemistry and biology

- Statistical Mechanics at atomic resolution, structure *and* dynamics:

  There was a sense, even at the time, of something truly historic going on, of getting these first glimpses of how an enzyme molecule for example, might undergo internal motions that allow it to function as a biological catalyst. (J.A.McCammon, Oral History (1995))

  $\Rightarrow$ MD as "computational microscope" (e.g., see here and here)

  MD generates a huge amount of data — "needle in the haystack"

MD in computational chemistry and biology

- Statistical Mechanics at atomic resolution, structure *and* dynamics:

  There was a sense, even at the time, of something truly historic going on, of getting these first glimpses of how an enzyme molecule for example, might undergo internal motions that allow it to function as a biological catalyst. (J.A.McCammon, Oral History (1995))

  $\Rightarrow$ MD as "computational microscope" (e.g., see here and here)

  MD generates a huge amount of data — "needle in the haystack"

- Building block for further applications (e.g., "free energy simulations")

Molecular dynamics (MD) in a nutshell

One particle

force=mass×acceleration

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

i.e.

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

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

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 2^{nd} order differential equation (Initial condition: $\mathbf{r}$ and $\mathbf{v}$ at $t = 0$)

N particles

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

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

$$\ldots$$

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

$\Rightarrow$ Numerical integration of the equations of motion

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 2^{nd} order differential equation (Initial condition: $\mathbf{r}$ and $\mathbf{v}$ at $t = 0$)

N particles

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

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

$$\ldots$$

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

$\Rightarrow$ Numerical integration of the equations of motion

Readily available tools

- Force fields
- Programs
- Tools for setup

Key ingredients to meaningful MD simulations

- Accurate force field
- Sufficient sampling
- Correct system preparation and setup

Key ingredients to meaningful MD simulations

We are restricted to toy models of reality

- force fields are approximate
- Limits to system size/composition and simulation length

$\Rightarrow$ System preparation and setup is crucial!

Key ingredients to meaningful MD simulations

We are restricted to toy models of reality

- force fields are approximate
- Limits to system size/composition and simulation length

$\Rightarrow$ System preparation and setup is crucial!

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

Force fields & Sampling

Force fields

- AMBER, CHARMM, GROMOS, OPLS-AA/M, Open Force Field Initiative, OPLS3e (Schrödinger)
- Proteins, DNA/RNA, fatty acids, membranes, carbohydrates, drug-like small molecules, modifications of amino acids etc.

Force fields & Sampling

Force fields

- AMBER, CHARMM, GROMOS, OPLS-AA/M, Open Force Field Initiative, OPLS3e (Schrödinger)
- Proteins, DNA/RNA, fatty acids, membranes, carbohydrates, drug-like small molecules, modifications of amino acids etc.
- Most simulation programs support more than one of the above force fields
- The different force field "families" are not fully compatible (do not "mix and match"); use the tools for "your" force field to generate missing parameters

Force fields & Sampling

Force fields

- AMBER, CHARMM, GROMOS, OPLS-AA/M, Open Force Field Initiative, OPLS3e (Schrödinger)
- Proteins, DNA/RNA, fatty acids, membranes, carbohydrates, drug-like small molecules, modifications of amino acids etc.
- Most simulation programs support more than one of the above force fields
- The different force field "families" are not fully compatible (do not "mix and match"); use the tools for "your" force field to generate missing parameters

Sampling

- Repeat simulations
- Multiple shorter simulations "better" than one long simulation (at least most of the time)

Building a system for MD

- Get (reasonable) starting coordinates
- Deal with missing coordinates
- Put biomolecule in water or membrane, etc.
- Add other molecules, components if needed
- Reflect experimental conditions of your simulation system
  - Protonation states
  - What ion types to use
  - Membrane Composition
  - Phosphorylation
  - Glycosylation
  - ...

Building a system for MD

- Get (reasonable) starting coordinates
- Deal with missing coordinates
- Put biomolecule in water or membrane, etc.
- Add other molecules, components if needed
- Reflect experimental conditions of your simulation system
  - Protonation states
  - What ion types to use
  - Membrane Composition
  - Phosphorylation
  - Glycosylation
  - ...

Learn as much as possible about your system!

Sources of coordinates

- X-ray, NMR
- Cryo-electron microscopy (see, e.g., here)
- Integrative/hybrid (I/H) methods
- Homology modeling
- AlphaFold
- Assembling a larger structure from "bits and pieces"

Sources of coordinates

- X-ray, NMR
- Cryo-electron microscopy (see, e.g., here)
- Integrative/hybrid (I/H) methods
- Homology modeling
- AlphaFold
- Assembling a larger structure from "bits and pieces"

The "stability" of your simulation is correlated to the "quality" of your starting coordinates

Missing coordinates

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

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

Missing coordinates

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

- 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!)

Protonation (+ tautomeric state)

- Proteins: PROPKA (& PDB2PQR)
- Organic molecules:
  - Various (empirical) tools, e.g., ChemAxon, OpenEye, Epik (Schrödinger), ACD/pK_a, S+pKa ...
  - Fast QM based methods, e.g., JPC A 2017, 121, 699
- 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. $\Rightarrow$ Constant pH methods

Setting up the system

- Focus on important region(s)

Setting up the system

- Focus on important region(s)
- **Use tools / standard workflows**
  - Tools/workflows coming with the biomolecular MD program, QuikMD, Packmol,...

Setting up the system

- Focus on important region(s)
- **Use tools / standard workflows**
  - Tools/workflows coming with the biomolecular MD program, QuikMD, Packmol,...
  - Commercial tools: Maestro, MOE

Setting up the system

- Focus on important region(s)
- **Use tools / standard workflows**
  - Tools/workflows coming with the biomolecular MD program, QuikMD, Packmol,...
  - Commercial tools: Maestro, MOE
  - **CHARMM-GUI**
    See here for recordings and slides from the CECAM CHARMM-GUI School!! (go to the 'Documents' tab)

Setting up the system

- Focus on important region(s)
- **Use tools / standard workflows**
  - Tools/workflows coming with the biomolecular MD program, QuikMD, Packmol,...
  - Commercial tools: Maestro, MOE
  - **CHARMM-GUI**
    See here for recordings and slides from the CECAM CHARMM-GUI School!! (go to the 'Documents' tab)
- Document why you chose settings — if things go wrong, revisit those choices!
- Think in advance how to "gauge" the stability of your simulation — this depends on the complexity of the system you are setting up!

A look ahead

It's tough to make predictions, especially about the future.

Si tacuisses, philosophus mansisses

...

Force fields . . .

- Majority of biomolecular MD simulations uses FFs with functional form of Eq. 27 of **Lifson** and Warshel, JCP 49, 5116 (1968) — strictly additive interactions, point charges

Force fields . . .

- Majority of biomolecular MD simulations uses FFs with functional form of Eq. 27 of **Lifson** and Warshel, JCP 49, 5116 (1968) — strictly additive interactions, point charges
- Polarizable force fields are coming of age, e.g.,
  - AMOEBA
  - CHARMM Drude FF family
  - . . .
- By now, various GPU accelerated codes available for these FFs, computational cost for protein–ligand complexes is becoming acceptable. (E.g., using OpenMM, the cost for the Drude FF is 4$\times$ the cost of the analogous additive FF)

Force fields . . .

- Majority of biomolecular MD simulations uses FFs with functional form of Eq. 27 of **Lifson** and Warshel, JCP 49, 5116 (1968) — strictly additive interactions, point charges
- Polarizable force fields are coming of age, e.g.,
  - AMOEBA
  - CHARMM Drude FF family
  - . . .
- By now, various GPU accelerated codes available for these FFs, computational cost for protein–ligand complexes is becoming acceptable. (E.g., using OpenMM, the cost for the Drude FF is 4$\times$ the cost of the analogous additive FF)

Multi-scale models (2013 Nobel Prize!)

- QM/MM
- ??? MM + Coarse-grained models

Ever more complex systems

- ribosome, cellular motors, entire cellular compartments, molecular crowding

Ever more complex systems

- ribosome, cellular motors, entire cellular compartments, molecular crowding
- Recent examples: **SARS-CoV-2 Spike Protein**
  - *Structure, Dynamics, Receptor Binding, and Antibody Binding of Fully-glycosylated Full-length SARS-CoV-2 Spike Protein in a Viral Membrane* (set-up completely with CHARMM-GUI)(*)
  - *Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein*
  - *AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics*

Ever more complex systems

- ribosome, cellular motors, entire cellular compartments, molecular crowding
- Recent examples: **SARS-CoV-2 Spike Protein**
  - *Structure, Dynamics, Receptor Binding, and Antibody Binding of Fully-glycosylated Full-length SARS-CoV-2 Spike Protein in a Viral Membrane* (set-up completely with CHARMM-GUI)(*)
  - *Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein*
  - *AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics*

(*) Building, setting up this system took 2 months!

Ever more complex systems

- ribosome, cellular motors, entire cellular compartments, molecular crowding
- Recent examples: **SARS-CoV-2 Spike Protein**
  - *Structure, Dynamics, Receptor Binding, and Antibody Binding of Fully-glycosylated Full-length SARS-CoV-2 Spike Protein in a Viral Membrane* (set-up completely with CHARMM-GUI)(*)
  - *Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein*
  - *AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics*

  (*) Building, setting up this system took 2 months!
- Need for and use of enhanced sampling, methods like Markov chain models (see here, here, and here), and other tools building on MD ($\Rightarrow$ alchemical FES) will increase

AI, neural nets, deep learning, etc.

- AlphaFold — breakthrough in the structure prediction of globular proteins (paper, database and a do it yourself Colab notebook)

AI, neural nets, deep learning, etc.

- AlphaFold — breakthrough in the structure prediction of globular proteins (paper, database and a do it yourself Colab notebook)
- (Q(uantum)) M(achine) L(earning): e.g., PhysNet, ANI, and link to a recent review (Ann.Rev.Phys.Chem. 2020, 71, 361). High level quantum chemistry at affordable cost, but see, e.g., an application to solute–solvent systems

AI, neural nets, deep learning, etc.

- AlphaFold — breakthrough in the structure prediction of globular proteins (paper, database and a do it yourself Colab notebook)
- (Q(uantum)) M(achine) L(earning): e.g., PhysNet, ANI, and link to a recent review (Ann.Rev.Phys.Chem. 2020, 71, 361). High level quantum chemistry at affordable cost, but see, e.g., an application to solute–solvent systems
- Optimization of traditional force fields (better combination of raw data, deriving FFs "from scratch" and use of QML to optimize dihedral terms)

AI, neural nets, deep learning, etc.

- AlphaFold — breakthrough in the structure prediction of globular proteins (paper, database and a do it yourself Colab notebook)
- (Q(uantum)) M(achine) L(earning): e.g., PhysNet, ANI, and link to a recent review (Ann.Rev.Phys.Chem. 2020, 71, 361). High level quantum chemistry at affordable cost, but see, e.g., an application to solute–solvent systems
- Optimization of traditional force fields (better combination of raw data, deriving FFs "from scratch" and use of QML to optimize dihedral terms)
- Analysis: Detect rare events, relevant degrees of freedom (one example, SARS-CoV-2 spike protein, full SARS-CoV-2 viral envelope) $\Leftrightarrow$ statistical learning

AI, neural nets, deep learning, etc.

- AlphaFold — breakthrough in the structure prediction of globular proteins (paper, database and a do it yourself Colab notebook)
- (Q(uantum)) M(achine) L(earning): e.g., PhysNet, ANI, and link to a recent review (Ann.Rev.Phys.Chem. 2020, 71, 361). High level quantum chemistry at affordable cost, but see, e.g., an application to solute–solvent systems
- Optimization of traditional force fields (better combination of raw data, deriving FFs "from scratch" and use of QML to optimize dihedral terms)
- Analysis: Detect rare events, relevant degrees of freedom (one example, SARS-CoV-2 spike protein, full SARS-CoV-2 viral envelope) $\Leftrightarrow$ statistical learning

Challenges, questions?

- Where is the physics?
- Interaction with water, environment? (Towards ML implicit solvation models)
- AI/MM (one example)

Concluding remarks

- The difficulty/challenge today is setting up a meaningful model; running a simulation is (relatively easy)
- Tools (e.g., CHARMM-GUI) to the rescue!
- (Biochemical/biological) domain knowledge will become (even more) important
- Computers will continue to get faster; how to use this power is limited by our imagination and ingenuity