Some mathematical challenges in molecular dynamics

Gabriel STOLTZ

(CERMICS, Ecole des Ponts & MATHERIALS team, INRIA Paris)

Work also supported by ANR Funding ANR-14-CE23-0012 ("COSMOS")

Presentation of the institutions at play

- **CERMICS: Applied mathematics laboratory of Ecole des Ponts**
    - 18 permanent members (16 HdR)
    - about 30 PhD students and 10 postdocs
    - research directions:
        - applied probability
        - modeling, analysis and simulation
        - optimization and operations research

- **MATHERIALS: project-team of Inria Paris**
    - 8 permanent researchers
    - strong overlap with CERMICS
    - analysis and development of simulation methods for multiscale models (incl. stochastic homogenization) and molecular simulations (quantum/classical)

What can (and cannot) applied mathematics do?

We, applied mathematicians...

- work on simplified models (one dimensional reaction coordinates, overdamped Langevin dynamics, etc) → this make us look like fools...
- exaggerate sources of errors (for instance by considering situations which rarely happen in practice) → this makes us insufferable

...but the aim is to

- rigorously understand why some methods work and some don't (mathematical proofs)
- devise new numerical strategies based on this theoretical understanding
- validate them on toy examples
- transfer the knowledge to practitionners by helping them to implement the methods into their own codes

Computational Statistical Physics

- *Predict macroscopic properties of materials from their microscopic description*

- **Microstate**
  - positions $q = (q_1, \ldots, q_N)$ and momenta $p = (p_1, \ldots, p_N)$
  - energy of the system $H(q, p) = V(q) + \sum_{i=1}^{N} \dfrac{p_i^2}{2m_i}$
  - (almost) all the physics is in the choice of V ...

- **Macrostate**
  - described by a probability measure μ
  - constraints fixed exactly or in average (number of particles, volume, energy)

- **Properties:** static (equation of state, heat capacities, etc) and dynamical (transport coefficient, transition pathway, etc)
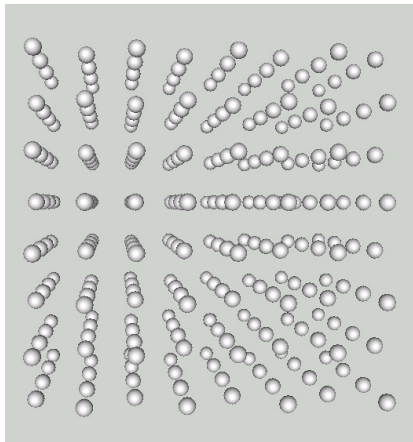
Examples of molecular systems (1)
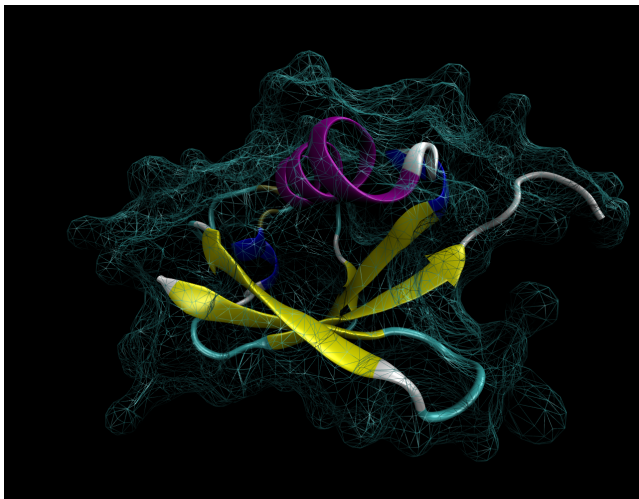
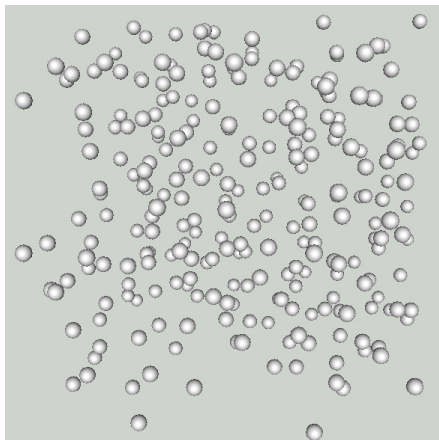**Ubiquitin (protein):** structure? conformational changes?
$\rightarrow$ In silico drug design

Examples of molecular systems (2)

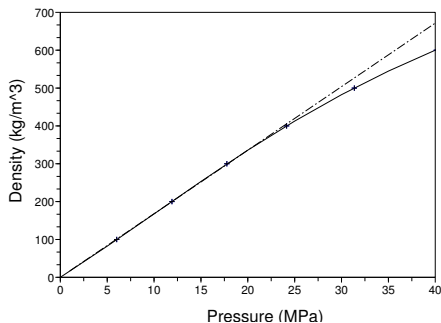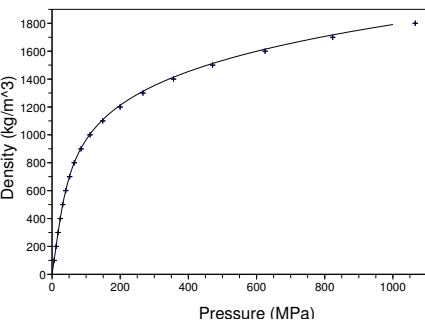What is the melting temperature of Argon?

(a) Solid Argon (low temperature)

(b) Liquid Argon (high temperature)

Examples of molecular systems (3)

Equation of state of Argon: density as a function of pressure, $T = 300$ K
(comparison with data of *National Institute of Standards and Technology*)

Explore extreme conditions of matter...

Some orders of magnitude...

- **Physical quantities**
  - distances $\sim 1 \ \mathring{A} = 10^{-10}$ m
  - energy per particle $\sim k_{\mathrm{B}}T \sim 4 \times 10^{-21}$ J at 300 K
  - atomic masses $\sim 10^{-26}$ kg
  - typical times $\sim 10^{-15}$ s
  - number of particles $\sim \mathcal{N}_A = 6.02 \times 10^{23}$

- **"Standard" simulations**
  - 10^6 particles ["heroic": 10^9 particles and more]
  - total time: (fraction of) ns ["heroic": (fraction of) μs]

- **Analogy** to understand what such large numbers represent...
  - about 10^{22} moles of water on Earth
  - 10^6 moles of water ~ 1 m^3

Aims of computational statistical physics

- **"Numerical microscope"**
  - gaining some insight into physical mechanisms at the atomic scale
  - From the press release for the Nobel prize in Chemistry 2013 (Karplus/Levitt/Warshel)

    Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments.

- **Computation of average properties** (static)

$$\langle A \rangle = \int_{\mathcal{E}} A(q, p) \, \mu(dq \, dp)$$

Computation of high dimensional integrals

  - μ is a probability measure
  - A is the observable

Outline

- **Computation of average properties** (free energy, error estimation)

- Computation of **transport coefficients**

- Sampling **reactive trajectories**

Main message

A careful construction of the numerical method can reduce both systematic and statistical errors

Note: takes a lot of time to **understand hence improve** numerical methods in molecular dynamics [and the impetus comes from various channels...]

T. Lelièvre and G. Stoltz, Partial differential equations and stochastic methods in molecular dynamics, *Acta Numerica* **25**, 681-880 (2016)

Computation of

average properties

Free energy computations (1)

- Canonical distribution $\mu(dq\,dp) = Z^{-1}\mathrm{e}^{-\beta H(q,p)}\,dq\,dp$

- For a given reaction coordinate $\xi(q)$, compute the function

$$F(z) = \int_{\xi^{-1}\{z\}} \mathrm{e}^{-\beta H(q,p)}\,\delta_{\xi(q)-z}(dq)$$

Thermodynamic integration, free energy perturbation, nonequilibrium techniques, adaptive methods (ABF, $\star$-metadynamics, ...)

- ABF in the simplest case $(\xi(q) = q_1)$

$$\left\{ \begin{array}{l} dq_t = -\nabla\Big(V - F_t \circ \xi\Big)(q_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t, \\ F_t'(z) = \mathbb{E}\left(f(q_t)\,\Big|\,\xi(q_t) = z\right) \end{array} \right\}$$

Free energy computations (2)

- **Our contributions for ABF include:**
  - a theoretical understanding of the improved convergence rate compared to unbiased dynamics (entropy methods)
  - improvements of the simulation methods: use of replicas and selection (implemented in NAMD)
  - multidimensional reaction coordinate case: projecting the current estimated mean force onto a gradient

- **Other results for adaptive dynamics:**
  - convergence of Wang-Landau type dynamics (including Self-Healing Umbrella Sampling and Well-Tempered metadynamics)
  - suggestion of modifications in the methods to improve convergence rates (e.g. allow for larger steps in the free energy update for WTM)

- **Also results for other types of dynamics/methods:**
  - constraints (overdamped/underdamped, TI, Jarzynski–Crooks, etc)
  - temperature accelerated molecular dynamics

Finding good reaction coordinates

- Reaction coordinates are often chosen based on chemical intuition

- Use machine learning techniques to automatically find RC?
  - supervised techniques (known metastable conformations)
  - unsupervised techniques (autoencoders, tICA/EDMD, etc)

PhD thesis started in November in collaboration with Sanofi

Errors on average properties (1)

- Errors due to **timestep discretization**?

- Positions $q \in \mathcal{D} = (L\mathbb{T})^d$ or $\mathbb{R}^d$ and momenta $p \in \mathbb{R}^d$
  $\rightarrow$ phase-space $\mathscr{E} = \mathcal{D} \times \mathbb{R}^d$

- Hamiltonian (more general kinetic energies $U(p)$ can be considered[1])

$$H(q, p) = V(q) + \frac{1}{2} p^T M^{-1} p$$

Stochastic perturbation of the Hamiltonian dynamics

$$\begin{cases} dq_t = M^{-1} p_t \, dt \\ dp_t = -\nabla V(q_t) \, dt - \gamma M^{-1} p_t \, dt + \sqrt{\dfrac{2\gamma}{\beta}} \, dW_t \end{cases}$$

- Friction $\gamma > 0$ and inverse temperature $\beta = \dfrac{1}{k_B T}$

---

[1]Redon, Stoltz, Trstanova, *J. Stat. Phys.* (2016)

Errors on average properties (2)

- Decompose the dynamics in three parts as

$$(A) \quad \begin{cases} dq_t = M^{-1}p_t \, dt, \\ dp_t = 0, \end{cases}$$

$$(B) \quad \begin{cases} dq_t = 0, \\ dp_t = -\nabla V(q_t) \, dt, \end{cases}$$

$$(C) \quad \begin{cases} dq_t = 0, \\ dp_t = -M^{-1}p_t \, dt + \sqrt{\dfrac{2}{\beta}} \, dW_t. \end{cases}$$

- First order splitting schemes: Trotter splitting

$$P_{\Delta t}^{ZYX} = e^{\Delta t Z} e^{\Delta t Y} e^{\Delta t X} \simeq e^{\Delta t (A+B+\gamma C)}$$

- Second order schemes: Strang splitting

$$P_{\Delta t}^{ZYXYZ} = e^{\Delta t Z/2} e^{\Delta t Y/2} e^{\Delta t X} e^{\Delta t Y/2} e^{\Delta t Z/2}$$

- Other category: Geometric Langevin algorithms, *e.g.* $P_{\Delta t}^{\gamma C,A,B,A}$

Errors on average properties (3)

- $P_{\Delta t}^{B, A, \gamma C}$ corresponds to
$$\begin{cases} \widetilde{p}^{n+1} = p^n - \nabla V(q^n) \, \Delta t, \\ q^{n+1} = q^n + \Delta t \, M^{-1} \widetilde{p}^{n+1}, \\ p^{n+1} = \alpha_{\Delta t} \widetilde{p}^{n+1} + \sqrt{\dfrac{1 - \alpha_{\Delta t}^2}{\beta} M} \, G^n \end{cases}$$

where G^n are i.i.d. Gaussian and $\alpha_{\Delta t} = \exp(-\gamma M^{-1} \Delta t)$

- $P_{\Delta t}^{\gamma C, B_\eta, A, B_\eta, \gamma C}$ for
$$\begin{cases} \widetilde{p}^{n+1/2} = \alpha_{\Delta t/2} p^n + \sqrt{\dfrac{1 - \alpha_{\Delta t}}{\beta} M} \, G^n, \\ p^{n+1/2} = \widetilde{p}^{n+1/2} - \dfrac{\Delta t}{2} \nabla V(q^n), \\ q^{n+1} = q^n + \Delta t \, M^{-1} p^{n+1/2}, \\ \widetilde{p}^{n+1} = p^{n+1/2} - \dfrac{\Delta t}{2} \nabla V(q^{n+1}), \\ p^{n+1} = \alpha_{\Delta t/2} \widetilde{p}^{n+1} + \sqrt{\dfrac{1 - \alpha_{\Delta t}}{\beta} M} \, G^{n+1/2} \end{cases}$$

Errors on average properties (4)

• Invariant measure $\mu_{\gamma,\Delta t}$ different from μ

Error estimates on the bias

$$\int_{\mathscr{E}} \varphi \, d\mu_{\gamma,\Delta t} = \int_{\mathscr{E}} \varphi \Big(1 + \Delta t^p f_\gamma \Big) d\mu + \mathrm{O}(\Delta t^{p+1})$$

with

- $p = 1$ for Trotter splitting
- $p = 2$ for Strang splitting and GLA

• **Various asymptotic results** in the Hamiltonian limit $\gamma \to 0$ or in the overdamped limit $\gamma \to +\infty$

$\to$ e.g. understanding why the BAOAB scheme is much better than other discretizations of Brownian dynamics

B. Leimkuhler, Ch. Matthews and G. Stoltz, The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics, *IMA J. Numer. Anal.* **36**(1), 13-79 (2016)

Computation of
transport coefficients

Computation of transport coefficients

- Thermal conductivity, shear viscosity, etc: takes a lot of time to estimate (possibly **several months**)

- **Green-Kubo formulas:** integrated correlation functions

Effective diffusion at equilibrium

Unperiodized displacement $Q_t - Q_0 = \displaystyle\int_0^t M^{-1} p_s \, ds$

$$D = \lim_{t \to +\infty} \frac{\mathbb{E}\left[(Q_t - Q_0) \otimes (Q_t - Q_0)\right]}{2t} = \int_0^{+\infty} \mathbb{E}_0\left[M^{-1} p_t \otimes M^{-1} p_0\right] dt$$

- Alternatively: linear response of steady-state nonequilibrium dynamics

- **Issues/questions :**
  - bias due to discretization in time
  - variance reduction (in progress)

Results in 1D for $\varphi = \psi = V'$ and cosine potential
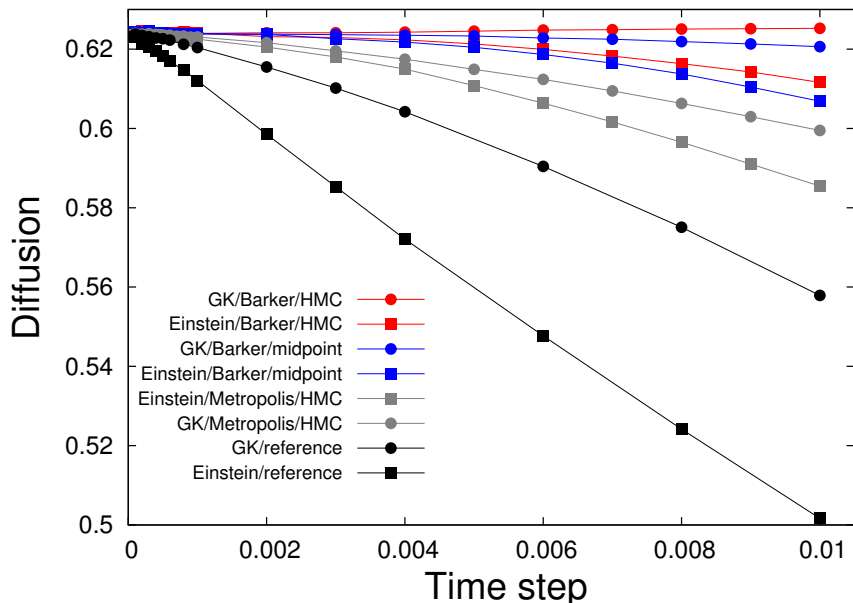
Sampling

reactive trajectories

Sampling reactive trajectories (1)

- Motivation: Unbinding of a ligand from a protein

Trypsin with various conformational states of benzamidine
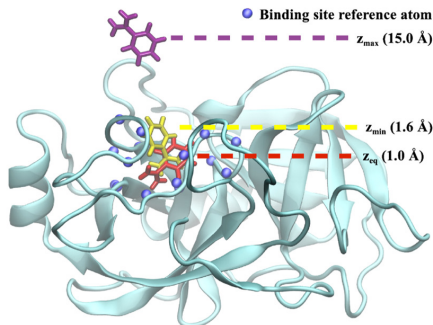
- Challenge: bridge the gap between timescales
  - Elementary time-step for the molecular dynamics $= 10^{-15}\,\mathrm{s}$
  - Dissociation time $\simeq 0.02\,\mathrm{s}$

Sampling reactive trajectories (2)

- **Mathematical setting: rare event computation**
  - Stochastic process $(X_t)_{t \geq 0}$, stopping times τ_A and τ_B
  - Aim: simulate and compute the very small probability $\mathbb{P}(\tau_B < \tau_A)$
  - Here sets A and B defined as metastable states (bound/unbound)

- Splitting technique: find intermediate events easier to simulate

$$\{\tau_{z_1} < \tau_A\} \supset \{\tau_{z_2} < \tau_A\} \supset \ldots \supset \{\tau_{z_{\max}} < \tau_A\} \supset \{\tau_B < \tau_A\}$$

and simulate the successive conditional events: for $k = 1, 2, \ldots$,

$$\{\tau_{z_q} < \tau_A\} \text{ knowing that } \{\tau_{z_{q-1}} < \tau_A\}$$

where $\tau_z = \inf\{t, \xi(X_t) > z\}$ for good importance function $\xi \in \mathbb{R}$.

- Adaptive feature: build the intermediate levels $(z_i)_{i \geq 1}$ on the fly $\rightarrow$ Adaptive Multilevel Splitting algorithm (Cérou/Guyader, 2007)
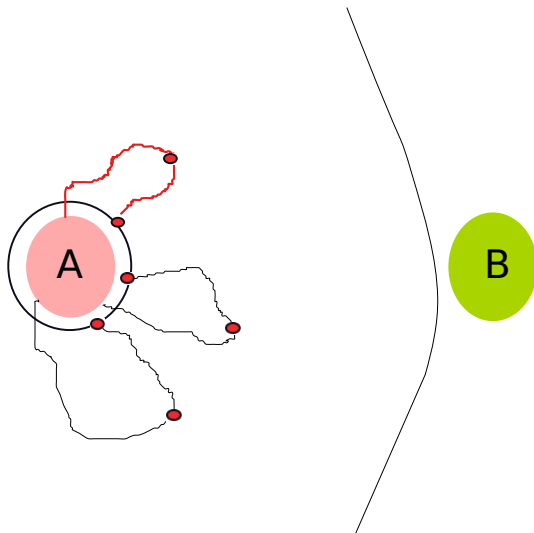
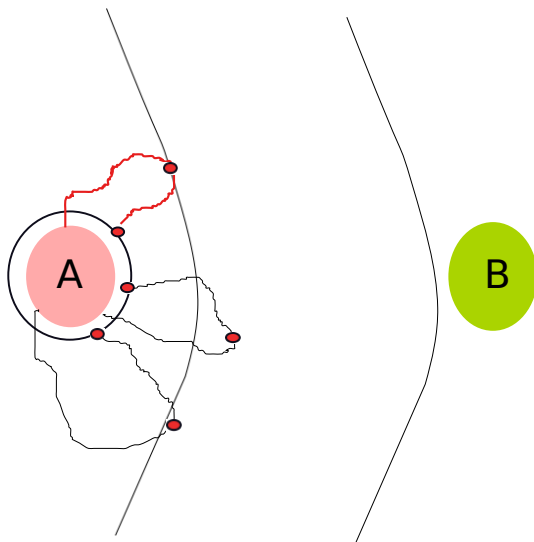# Schematic illusration of the AMS Algorithm
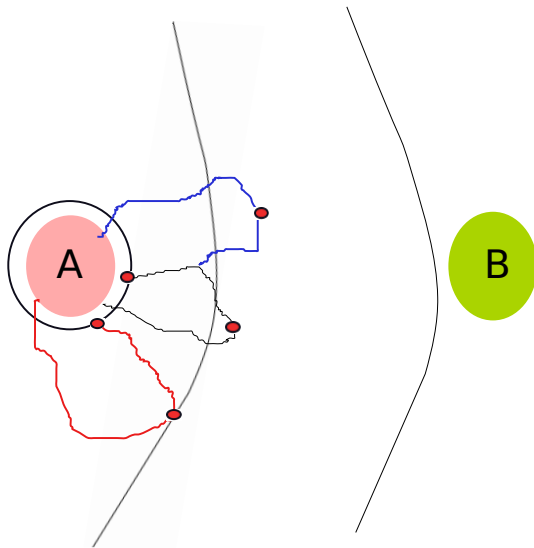
Sampling reactive trajectories (3)

• In collaboration with the group of K. Schulten (C. Mayne and I. Teo), AMS is currently implemented in the NAMD code.

• **Unbinding event of benzamidine from trypsin:**
  - MD setup: about 70 000 atoms, CHARMM36 force field, NPT conditions (298 K)
  - Estimated dissociation rate: $k_{\text{off}} = (260 \pm 240)s^{-1}$ which is in the same order of magnitude as the experimental rate $(600 \pm 300)s^{-1}$
  - Overall simulation time: $2.3 \ \mu s$ which is 4 orders of magnitude shorter than than the estimated dissociation time

T. Lelièvre and L. Lopes, Analysis of the Adaptive Multilevel Splitting method with the alanine di-peptide's isomerization, *Journal of Computational Chemistry*, **40**(11), 1198-1208 (2019)