PhD thesis proposal: *Machine learning for sampling complex biological systems, and vice versa*

- **Supervisors:** Pierre Monmarché (LJLL et LCT, Sorbonne Université)
  pierre.monmarche@sorbonne-universite.fr
  https://www.ljll.fr/~monmarche
  - Jérôme Hénin (LBT, IBPC, CNRS)
    jerome.henin@cnrs.fr
    http://www-lbt.ibpc.fr/people/henin

- **Host lab:** Laboratoire Jacques-Louis Lions (Sorbonne Université)
- **Domain:** interface between Mathematics, Machine Learning, Statistical Physics, and Computational Chemistry
- **Starting date:** October 2024

Candidate profile and prerequisites:

This is an interdisciplinary project, since Pierre Monmarché is a mathematician specializing in stochastic algorithms in high dimension, and Jérôme Hénin is a chemical biophysicist specializing in the development of free energy methods for molecular biophysics. The PhD project is open to candidates with different profiles:

- applied mathematics
- computer science for machine learning
- computational chemistry or physics

The work will be developed with more or less emphasis in its various aspects, depending on the doctoral student. The student is not expected to have mastered all these fields before the thesis.

Required skills:
- probability or statistical physics
- numerical analysis
- Python programming

Useful skills:
- use of ML frameworks (scikit-learn, pytorch)
- numerical simulation tools
- C++ programming

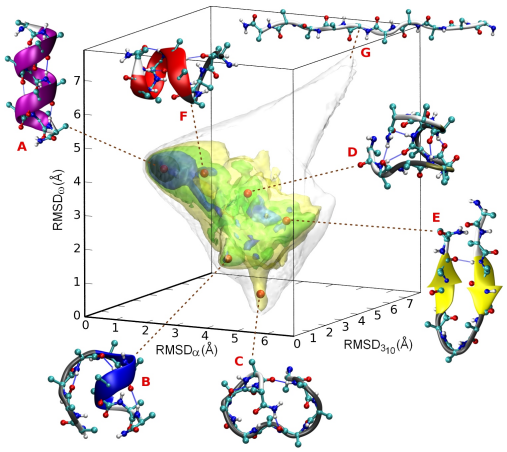
Scientific project:

One of the main difficulties in simulating molecular systems, particularly biological systems (such as proteins, made up of hundreds of thousands of atoms and with a complex, constrained geometry) is sampling. When naively simulating atom dynamics, the conformational changes of interest are prohibitively long and cannot be observed (the process is metastable: it remains stuck in certain regions for a long time before escaping).

Nowadays, high-performance enhanced sampling algorithms are available to bias the dynamics in such a way as to observe these transitions in practice. These methods (metadynamics, Adaptive Biasing Force...) are based on a principle of **dimension reduction**: the idea that, although described in high dimension, the system essentially evolves on a low-dimensional manifold. In particular, these methods require reaction coordinates, i.e. a very small number of variables that parametrize the low-dimension manifold. Statistical learning methods are increasingly used to obtain these variables. This is the background to the present research project, which focuses more specifically on the following three points:

1) Estimating the free energy in high dimension. Current adaptive bias methods [1] are limited to a small number of reaction coordinates (one or two in most cases), due to the curse of dimen-

sionality (the bias, a function of the reaction coordinates, is often stored on a grid). However, the physically relevant domain containing the system's probable configurations may be of locally greater dimension, meaning that current sampling methods are unable to explore it efficiently. The advent of unsupervised learning of reaction coordinates can easily produce dozens of them, so we need to overcome this limitation of sampling methods. This can be done by using statistical inference methods for bias and parsimonious nonlinear approximations (tensors, for example) [2], or by kernel methods coupled with local dimension estimates [3], but these methods need further development to bring them to the level of concrete applications.



2) Sampling in high dimension. In dimension 1 or 2, current methods seek to sample the uniform distribution, i.e. to visit all possible values of the reaction coordinates indiscriminately. However, uniformly exploring a hypercube of dimension 10 doesn't work in practice (the stochastic process wanders without finding the areas of interest). Appropriate methods need to be developed, analyzed and implemented, associating each reaction coordinate with an effective sampling temperature (the most important coordinates being explored uniformly, the less important coordinates being explored at a temperature higher than the initial unbiased system, but finite), parametrized adaptively (based on criteria such as the Poincaré constant of the local equilibrium measure). Another possible direction is to exploit the dynamics of saddle point search on the free energy surface [4].

3) Exploring the parameter space of Deep Learning. An important consequence of automated learning of reaction coordinates is that we can now apply adaptive bias methods to high-dimensional systems for which we have no physical intuition, such as the parameter space of artificial neural networks. It then becomes possible to identify key variables in these networks, calculate associated free energies to better understand their operation and the learning process (e.g. by observing the trajectory on this free energy landscape of stochastic gradient descents initialized with Gaussian variables, as is the practical use in learning) and improve optimization through local stochastic exploration of the parameters.

The aim of this thesis is to address these issues by developing high-performance codes that can be used by practitioners on real-life applications.

References

[1] **J. Hénin**, T. Lelièvre, M.R. Shirts, O. Valsson, L. Delemotte. *Enhanced sampling methods for molecular dynamics simulations*. LiveCoMS, 2022.

[2] V. Ehrlacher, T. Lelièvre, **P. Monmarché**. *Adaptive force biasing algorithms: new convergence results and tensor approximations of the bias*. AAP, 2022.

[3] A. Rodriguez, M. d'Errico, E. Facco, A. Laio. *Computing the free energy without collective variables*. JCTC, 2018.

[4] L. Journel, **P. Monmarché**. *Switched diffusion processes for non-convex optimization and saddle points search.* Statistics and Computing, 2023.