Sampling Molecular Distributions with Deep Neural Networks

Keywords: Deep Learning, Molecular Dynamics, Statistical Physics
Research team: TAU, Inria Saclay
(In collaboration with Laboratoire de Biochimie Théorique, CNRS, IBPC, Paris)
Location:
Laboratoire de Recherche en Informatique, Université Paris Sud
Bât 660 Claude Shannon, Rue Noetzlin, 91190 Gif-sur-Yvette, France
Supervisors:
Guillaume Charpiat (*guillaume.charpiat@inria.fr*) for the deep learning side
Jérôme Hénin (*jerome.henin@ibpc.fr*) for the bio-physical side

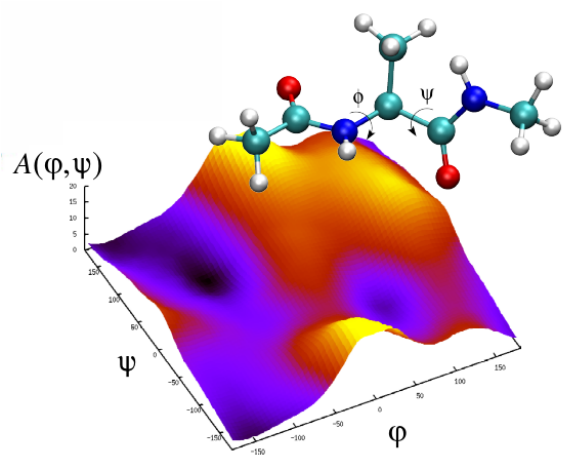

Figure 1: Distribution of *Alanine dipeptide* conformations, projected onto 2D relevant descriptors.

Scientific Context: Molecules (typically, proteins) do not have stable 3D configurations, they move very quickly between different conformations. Chemistry (in particular cellular biology) strongly depends on the **distribution of 3D conformations** the molecules undergo, which is thus of primal interest. While numerical simulations are used routinely to study the dynamics of biomolecules at the atomic scale, they produce highly correlated time series that may fail to sample the highly multimodal distribution of relevant molecular configurations. A powerful way to bypass this issue is to use a type of **generative neural networks** named **normalizing flows** that **map a multivariate Gaussian noise to the distribution of interest**. Building and training such networks pose majors challenges in terms of architecture, loss function design, and training strategies.

Research Context: One major obstacle faced by the current approaches to generate a 3D conformation distribution for a given molecule is that they depend on training data – typically obtained from a molecular simulation, which strongly limits the benefits of such generative approaches. Instead, it is possible to train models **without training data**, thanks to the evaluation of the generated samples through their **Boltzmann energy**. Unfortunately, in practice, for reasons not understood yet, current data-free approaches universally exhibit mode drop, preventing the generation of the correct multimodal distribution.

Objectives: The goal of this internship is to improve an existing proof-of-concept solution to this distribution generation task, in order to make it scale to larger problems. To that effect, the intern will work on adapting the architecture to improve its expressivity, or/and work on the task modelisation to reduce the need for expressivity:

- *Design a molecular modeling approach that makes generation easier.* The task of the main network can be made easier by reducing the dimension of the problem. Indeed, while a 3D conformation of a molecule with N atoms is of high dimensionality $3N$ (think of $N = 100$), it turns out that the "interesting" degrees of freedom are far fewer (like 4 or 5), and this brings the hope for an actually low dimension task. To achieve this, one needs to create a separate treatment for "degenerate" degrees of freedom, due to symmetries (translational and rotational invariance), and for very "local" degrees of freedom, not impacting the global structure of the molecule, such as atom vibrations or hydrogen atom locations.

- *Adapt architecture to improve expressivity.* Conversely, the architecture of the core network could be augmented with new types of layers that help it express multimodal distributions, such as a layer explicitly designed to map a unimodal input to a multimodal output. Another promising track is to design a structure dealing hierarchically with the degrees of freedom.

Infrastructure: We provide:
- PyTorch code to start from
- data in arbitrarily large quantities if needed
- GPU cluster
- and theoretical insights.

Requirements:
- Mathematics (functional analysis, linear algebra, calculus, probabilities)
- Good Python coding skills (and knowledge of its ecosystem of scientific libraries)
- Basic experience with Pytorch and/or Tensorflow and/or Chainer
- Interest in or knowledge of Deep Learning
- Fluency in english
- Basic knowledge of Organic Chemistry and/or Molecular Biology would also be helpful.