Collective Variable Learning in Molecular Simulations

Belkacemi, Z., [1,2,3] Gkeka, P., [2] Stoltz, G.,[3] Lelièvre, T. [3]
*Belkacemi Zineb, PhD Student
[1] zineb.belkacemi@enpc.fr, zineb.belkacemi@sanofi.com
[2] *Structure Design and Informatics, Sanofi R&D, Chilly-Mazarin, France*
[3] *CERMICS, Ecole Nationale des Ponts et Chaussées, Université Paris-Est, Champs sur Marne, France*

Molecular Dynamics (MD) simulations have proven to be a very useful complementary tool, and sometimes even an alternative to experiments. Despite their wide use to capture fast occurring phenomena (e.g. hydrogen bonds occurrence), there are still many cases where the time scales accessible to MD simulations are far smaller than the time scales needed for the observation of important conformational changes of the systems under study (e.g. protein folding). Many Enhanced Sampling methods have emerged to accelerate the observation of such changes, but these methods rely on the knowledge of low-dimensional slow degrees of freedom, i.e. Collective Variables (CV), that can only be intuited for small simple systems.

For larger and more complex dynamical systems, Machine Learning and Dimensionality Reduction techniques can be used for the automatic identification of Collective Variables (CVs). These methods span from basic linear Dimension Reduction algorithms, such as PCA (Principal Component Analysis) or tICA (time-lagged Independent Component Analysis), to complex non-linear or kernel-based methods, such as Diffusion Maps or AutoEncoders, but also include Supervised Learning and Feature Engineering Algorithms.

In this talk, I will present a general overview of these methods and compare them on the extensively studied example of Alanine Dipeptide.