Inria Sophia Antipolis
Group: Algorithms-Biology-Structure
Frederic.Cazals@inria.fr
Web: `http://team.inria.fr/abs`

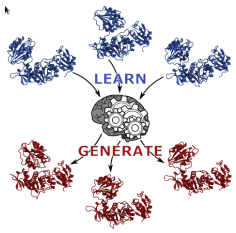

Figure 1: (Deep) Learning molecular properties: on which assets? Picture from [1].

Master internship proposal

Deep and wide: on the incidence of basin geometry in (deep) learning energy functionals

Keywords. High-dimensional spaces, thermodynamics, deep learning, energy landscapes, proteins.

Context. Macroscopic properties of biomolecules result from average properties computed over ensembles of conformations [2, 3]. More precisely, consider the mapping associating a potential energy $V(\cdot)$ to each conformation, the so-called *potential energy landscape* (PEL). PEL code all properties. The *structure* of a macromolecular system requires the characterization of conformations associated with significant (deep and/or wide) basins In assigning occupation probabilities to these conformations by integrating Boltzmann's distribution, one treats *thermodynamics*. Finally, transitions between the states correspond to *kinetics*. Since each atom has 3 Cartesian coordinates, a shear difficulty to study PEL is their huge dimensionality, namely $d = 3n(\gg 1000)$.

Because neural networks can learn real (vector) values smooth functions – cf. universal approximation theorems, it is natural to explore and compute average properties on PEL using deep learning.

Goals. Attempts have recently been made to discover i.e. map but also sample PEL using auto-encoders [4], and Boltzmann's samplers [5], namely machines trained on a given energy function $V(x)$ and then producing statistically independent samples from $exp(-V(x))$. At this stage, such results have a limited scope: they concern small systems, and most importantly, nothing is known on the correctness of the numerical quantities delivered.

On the other hand, the structure of an energy functional, say a PEL, can be summarized by a graph connecting significant local minima (whose basins are deep or wide) across saddle points [6, 7], and various geometric features summarizing the geometry of basins associated with local minima can be computed.

The goal of this internship will be to design and analyze provably correct deep learning algorithms computing certified approximations of ensemble averages on groups of basins. The developments will be validated on a set of models ranging from toy landscapes to PEL of biomolecules.

Conditions. Internship with *gratification*. Possibility to follow-up with a PhD thesis.

References

[1] M. Degiacomi. Coupling molecular dynamics and deep learning to mine protein conformational space. *Structure*, 27(6):1034–1040, 2019.

[2] D. J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

[3] D.M. Zuckerman. *Statistical Physics of Biomolecules: An Introduction*. CRC Press, 2010.

[4] Wei Chen and Andrew L Ferguson. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of computational chemistry*, 39(25):2079–2102, 2018.

[5] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[6] J. Carr, D. Mazauric, F. Cazals, and D. J. Wales. Energy landscapes and persistent minima. *The Journal of Chemical Physics*, 144(5):4, 2016.

[7] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert. Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J. Comp. Chem.*, 36(16):1213–1231, 2015.