Frédéric Cazals
Centre Inria at Université Côte d'Azur
Sophia-Antipolis
http://team.inria.fr/abs
Frederic.Cazals@inria.fr

Marylou Gabrié
Laboratoire de Physique, Ecole Normale Supérieure
Paris
https://marylou-gabrie.github.io/
marylou.gabrie@polytechnique.edu

Master internship proposal

Generative modeling under partial information: flowing efficiently and fast under oracle constraints

Keywords: generative models, molecules, conformational sampling, flat torii, diffusion models, flows, energy landscapes, min energy paths.

Context. Generating high quality and diverse molecular conformations of proteins is an open problem, whose difficulty owes to two main reasons: the non linear nature of molecular motions, and the high dimensionality of the systems studied ($d = 3n$, with n the number of atoms). Very recently, a number of generative methods based on transformers, normalizing flows and diffusion models have been proposed [1, 2, 3, 4, 5, 6]. The loss function driving the learning process typically uses all degrees of freedom, and requires high quality data, generally from molecular dynamics.

The goal of this internship is to bypass the need of large training sets, by trading them against a combo consisting of an explicit probabilistic model in a reduced space, and an implicit oracle in the full space.

Goals. Instead of using a global parameterization of the molecular system studied, assume we are given a joint probability distribution for selected degrees of freedom termed *reduced coordinates*, and an oracle returning a potential energy (or potential of mean force) associated with a conformation in reduced coordinates.

The goal of the internship is to develop a flow based generative method based on a variational problem encompassing two terms: the first akin to the classical one in (rectified) flow based methods; and the second one encoding the integral of some cost function.

Conditions. Internship with *gratification*.

Location: Warmup at ENS, then Inria Sophia-Antipolis with visits at ENS.

This internship may be followed by a PhD thesis.

References

[1] Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023. Cited 64 times as of Dec 2024.

[2] Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024. Cited 158 times as of Dec 2024.

[3] Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, pages 1–11, 2024. Cited 137 times as of Dec 2024.

[4] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024. Cited 42 times as of Dec 2024.

[5] Bowen Jing, Hannes Stark, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. In *NeurIPS*, 2024. Cited 4 as of Dec 2024.

[6] Simon Wagner, Leif Seute, Vsevolod Viliuga, Nicolas Wolf, Frauke Gräter, and Jan Stuehmer. Generating highly designable proteins with geometric algebra flow matching. In *NeurIPS*, 2024.