# ***INTERNSHIP PROPOSAL***

Laboratory name: Institut de Physique Theorique (IPhT)
CNRS identification code: UMR 3681
Internship director'surname: FERNANDEZ DE COSSIO DIAZ, Jorge
e-mail: jorge.fdcd@ipht.fr
Web page: https://sites.google.com/view/jorgefdcd
Internship location: IPhT, Université Paris-Saclay
Thesis possibility after internship: YES, subject to funding

## Manipulating specificity in biological sequences with representation learning

A biological sequence (DNA, RNA, protein) is a string of contiguous covalently attached amino acids or nucleotides. A central paradigm of biology is that the sequence determines the function of the molecule in the organism. However, this mapping is complex and context dependent. Understanding how widely diverse functions are encoded in these sequences is a fundamental question in biology. Sequence design can be viewed as the inverse map, from function to sequence: Given a desired phenotypic task, which sequences can perform it? Due to the complexity of the relationship between sequence and function, the design of biological sequences to achieve desired functions is a challenging problem, with numerous applications in drug development, industry, basic experimental research, synthetic biology, and others.

During evolution nature samples many sequence variants that perform closely related tasks in diverse organisms. Modern next generation sequencing gives access to large datasets, where sequence variability within families of homologous proteins and RNA can be inspected. The statistics of related sequences contain signatures about their evolutionary constraints.

Generative models trained on large sequence datasets can be used to sample novel functional sequences. However, generated sequences often reproduce statistics of the training data, combining various features found in the natural sequences in an uncontrolled manner.

In this internship we will modify generative models (such as RBM, VAE, Diffusion, …) to extract disentangled representations of biological sequences, where interesting properties are mapped to independent latent coordinates. Such latent variables can be modified during sampling to control specific features of designed sequences, without interfering with other properties. We are also interested in theoretical understanding of representation learning using analytical techniques from statistical physics approaches to neural networks.

Background (and interest) in topics at the intersection of statistical physics, machine learning, and biology, as well as programming experience (e.g. PyTorch, Julia, …) are strongly advised.

**References:** • JFdCD, et al. PRX 13.2 (2023): 021003. • JFdCD, et al. bioRxiv 2023.05.10.540155. • F. Locatello, et al. ICML, 2019. • G. Lample et al. NIPS 30 (2017). • JFdCD, et al. hal-04447899 (2024).

Please, indicate which speciality(ies) seem(s) to be more adapted to the subject:

| | |
|---|---|
| Condensed Matter Physics: YES | Soft Matter and Biological Physics: YES |
| Quantum Physics: NO | Theoretical Physics: YES |