Laboratory name: Institut de Physique Théorique (IPhT). CNRS id code: UMR 3681
Directors: J. FERNANDEZ DE COSSIO DIAZ & P. URBANI
e-mail: jorge.fdcd@ipht.fr, pierfrancesco.urbani@ipht.fr
web page: jorgefdcd, pierfrancesco-urbani
Thesis possibility after internship: Yes

Controlling and understanding representations in state-space models of biological sequences

State-space models process sequence data one token at a time. A state vector is updated with each consumed token, and stores a compressed "memory" of past tokens. To be effective, the model must be able to recognize and exploit long-range correlations between tokens along the sequence. This is a particularly important issue in applications of these models to biological sequences, such as proteins or RNA, where sites may be far-apart along the sequence, yet interact closely in the folded 3D structure of the molecule. Since the latent state vector encodes all the information that the model needs to construct future tokens of the sequence, one can try to find ways of controlling this latent vector during generation to modify aspects of the sequence, *e.g.*, to manipulate desired properties of an artificial protein.

We will attack these issues from both a computational and a theoretical side. On a computational side, we will apply state-space models (such as MAMBA [3]) to biological sequence data, and develop algorithms to manipulate state representation vectors to design sequences with desired properties (*e.g.* through disentangled representations [1] or guidance [5]). Designed molecules can be tested experimentally with collaborators [4]. On a theoretical side, we will formulate simplified models and exploit methods from statistical physics (e.g., dynamical mean-field theory [2]), aiming at understanding how these models learn about long-range interactions and how token representations can be controlled during generation to modify properties of designed sequences.

This interdisciplinary internship and PhD project are aimed at students with a strong background and interest in statistical physics, machine learning, biology, and coding.

References

[1] JFdCD, S. Cocco, R. Monasson. "Disentangling representations in restricted Boltzmann machines without adversaries." Physical Review X 13.2 (2023): 021003
[2] P.J. Kamali, PU. (2023) "Dynamical mean field theory for models of confluent tissues and beyond," *SciPost Physics*, 15(5), p. 219. Available at: https://doi.org/10.21468/SciPostPhys.15.5.219.
[3] D. Sgarbossa, et al. (2025) "ProtMamba: a homology-aware but alignment-free protein state space model," Bioinformatics, 41(6), p. btaf348.
[4] JFdCD, et al. "Designing molecular RNA switches with Restricted Boltzmann machines" Nat. Comm. (2025, in Press). bioRxiv 2023.05.10.540155.
[5] A. Bansal, et al. "Universal guidance for diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.