Postdoctoral position on explainable AI applied to genome data.

In the context of the GraphNEx European consortium involving Queen Mary University of London (UK), the Ecole Polytechnique Federal de Lausanne (EPFL, Switzerland) and the ENS de Lyon (France), we look for a motivated postdoctoral researcher to join ENS de Lyon as soon as possible and preferably prior to november 1st. The successful candidate will be associated in his research to Pierre Borgnat and Benjamin Audit both at the Physique Laboratory and to Jean-Michel Arbona at the Laboratoire de Biologie et Modelisation de la cellule.

Research context: In this project we propose to develop novel methods to explain the prediction of graph neural networks trained on biological data. The first part of the project will be to design an architecture and to train a neural on a large dataset of epigenetic data (100 G to 1 Tera octets). In the spirit of the AVOCADO (1) framework, the neural network will be designed as a combination of tensor factorization architecture and neural network. The combination of the two methods allows the development of a robust imputing system: while trained on partial data it is able to predict missing data. The main structure of the data follows the sequential nature of the DNA sequence while the observables are the epigenetic modifications associated to the sequence. One of the unknown and target of the project is the length of the sequence that is required to perform predictions at the middle of the DNA segment ranging from 1000 base pair to 100 kilo base pair. Of specific interest for us will be the prediction of the initiation landscape of the DNA replication process, for which only very partial data are available. The designed system will rely on the numerous datasets of epigenetics mark to improve our knowledge of replication.
A second part of the project will be to develop explainability methods to understand the prediction of the trained neural network. Ideally the explainability method will be focused on the graph part of the neural network architecture that in that context could be defined either as interaction of different parts of the DNA sequence, or on interaction between genes of a cellular type.

Background of the candidate :
The candidate will have experience in developing neural networks preferentially on sequence to sequence applications, or on graph data. Ideally the candidate will also have some familiarity with biological data.

(1) Schreiber, J., Durham, T., Bilmes, J., & Noble, W. S. (2020). Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome biology*, *21*(1), 1-18.