Surrogate network inference for heterogeneous data

Next generation sequencing technologies have given rise to a tsunami of biological data at unprecedented resolution, accuracy and scale. The extraordinary complexity of these data brings about immense challenges for mathematicians, and statisticians in particular. For instance, the question of jointly analyzing many variables of multiple types (continuous, ordinal, categorical) has attracted much attention recently, both in biology and statistics. This research question is driven by the need in biology to infer regulation networks from so called multi-omics data (representing different levels of molecular variability: genomics, ranscriptomics, proteomics, metabolomics, epigenomics, etc.) To "couple" distinct types of data, one approach is to build statistical models based on copulas parametrized by a correlation matrix. It is thus possible to identify independent groups of omics and discover approximate conditional independence relationships between them. The computational cost is, however, very large. It increases with the number of pairs of discrete variables, because for each of them a double integral has to be computed numerically. The objective of this postdoc is to address the above computational challenges. One possible research avenue is to replace the log-likelihood by a cheap surrogate learned on all possible data. The tradeoff between accuracy and computational complexity needs to be optimized both theoretically and numerically.
The candidate is expected to have a background in statistics or machine learning.