# PhD position in biostatistics and machine learning at CEA Grenoble

**Subject**: Calibration of protein sequence identification procedures for conformal predictions

**Key words**: metric learning; high dimensional statistics; conformal risk control

**Context**: The large-scale characterization of proteins in samples from living organisms (a.k.a, proteomics) requires analyzing short amino acid sequences (referred to as peptides) by a mass spectrometer. The resulting spectra are then matched onto a reference sequence database. When doing so, the discrepancies between the sequences and the spectra are generally quantified by scores of various complexity (ranging from a simple sequence coverage percentage to a mathematically well-defined positive and semi-definite metric). Unfortunately, these scores are difficult to interpret for biology researchers, which raises questions about how these peptide-to-spectrum matches (PSMs) can be further validated. Owing to the importance of this question, a multi-disciplinary consortium involving mass spectrometry experts, biologists, computer scientists and statisticians has been financed by the French National Research Agency.

**Missions**: The recruited PhD student will integrate our consortium to elaborate new methods to endow these PSM scores with well-defined statistical properties, so as to:

(i) Propose novel mathematically well-defined PSM metrics,

(ii) Estimate the risks associated to type I and type II errors,

(iii) Extend these risk control tools to multiple testing scenarios (i.e., when tens of thousands of PSMs are considered simultaneously);

(iv) Provide uncertainty estimates, notably when ambiguous spectra can be matches onto multiple different amino acid sequences.

A specific challenge of this project relies on the black-box nature of the engines used to generate the PSMs (either because these tools are commercial; or because the predictions are fueled by poorly interpretable neural networks). It will thus be necessary to leverage the recent advances in statistical learning that make it possible to derive asymptotic guarantees from black-bock machine learning-based decisions, like conformal predictions [1], prediction powered-inferences [2], conditional randomization testing procedures [3] and knock-off filters [4].

**Expected profile**:

- Master degree (or equivalent engineering degree) in statistics, in signal processing, in data science or in applied mathematics with advanced skills in scientific programming (R or Python).
- Software development skills in object-oriented language like Java is a bonus.
- Proficient either in French or in English.
- Motivated by the interdisciplinary nature of the project.

Applicants should send an extended version of their CV and an application letter to Thomas Burger (thomas.burger@cea.fr).

**References:**

[1] https://www.youtube.com/watch?v=nql000Lu_iE

[2] Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669-674.

[3] Liu, M., Katsevich, E., Janson, L., & Ramdas, A. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika*, *109*(2), 277-293.

[4] Candes, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(3), 551-577.