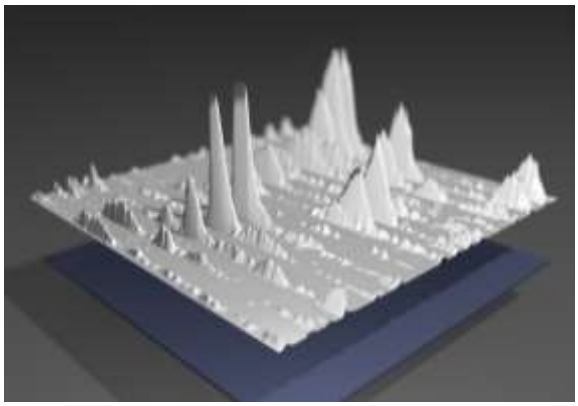
Machine learning based alignment of mass spectrometry data for improved clinical biomarker discovery



Context: Mass spectrometry (MS) is the method of choice to discover candidate biomarkers in clinical research. However, the increasing complexity of clinical samples combined to the stochastic non-exhaustive identifications yielded by MS, make it necessary to: (1) perform multiple analyses; (2) align them; (3) pool/transfer the identifications between analyses.

Challenges: To date, although few algorithms have been proposed to do so, there does not exist any generic methodological framework allowing for the alignment of several samples, in order to subsequently provide a robust identification transfer between samples. An innovation in this direction will therefore strongly impact the reproducibility of MS-based molecular profiling, and subsequently, clinical research tools.

Objectives: We propose to model the alignment of MS analyses by relying on multi-view learning techniques [1]. First of all, the metrics associated to misalignments caused by analytical biochemistry processing on the one hand, and those resulting from MS comparisons on the other hand, will be combined using two families of dedicated positive semi-definite kernels (in particular, kernels resulting from the theory of optimal transport will be investigated [2]). As, for dimensionality reasons, it is not possible to directly combine the kernels of the two views, an additional fusion stage, based on a semi-supervised strategy (*e.g.*, co-training [3]) will then be implemented.

Profile: Student in M2 or computer engineering school (Artificial Intelligence or Data Science specialty) who has a strong interest in interdisciplinary work. He/she must have programming skills (Python and Java) and be fluent in either French or English.

Hosting team: The PhD position will be hosted at CEA Grenoble, in EDyP team (www.edyp.fr) and supervised by Christophe Bruley (CEA engineer, head of EdyP, head developer of the Proline software suite, https://www.profiproteomics.fr/proline/, [4]) and Thomas Burger (CNRS senior scientist, https://sites.google.com/site/thomasburgerswebpage/). The PhD candidate will be affiliated to the EDISCE doctoral school (https://edisce.univ-grenoble-alpes.fr/) and will benefit from the fostering environment of the Interdisciplinary Institute for Artificial Intelligence (https://miai.univ-grenoble-alpes.fr/). Applicants should send their CV to C. Bruley and T. Burger (firstname.lastname@cea.fr), as well as other credentials (diploma, recommendation and motivation letters, etc.).

References:

[1] M. Gönen, E. Alpaydın. "Multiple kernel learning algorithms". The Journal of Machine Learning Research, 12, pp. 2211-2268, 2011.

[2] O. Permiakova, R. Guibert, A. Kraut, T. Fortin, A.-M. Hesse, T. Burger. "CHICKN: Extraction of peptide chromatographic elution profiles from large scale mass spectrometry data by means of Wasserstein compressive hierarchical cluster analysis". BMC Bioinformatics, 22.68, 2021.

[3] A. Blum, T. Mitchell. "Combining labeled and unlabeled data with co-training". In Proceedings of the eleventh annual conference on Computational learning theory, pp. 92-100, 1998.

[4] D. Bouyssié, A.-M. Hesse, E. Mouton-Barbosa, M. Rompais, C. Macron, C. Carapito, ..., C. Bruley. "Proline: an efficient and user-friendly software suite for large-scale proteomics". Bioinformatics, 36(10), pp 3148-3155, 2020.