Titouan Vayer (titouan.vayer@inria.fr)
Franck Picard (franck.picard@ens-lyon.fr) **October 30, 2025**

Intership proposal: Distributional perspective of self-supervised learning

Keywords: Self-supervised learning, optimal transport, dimension reduction, clustering.

Context

With the joint progress in collecting increasingly complex and massive datasets and the rapid rise of large foundation models, it has become crucial to design methods that can adapt to multiple tasks and data modalities. This raises a fundamental challenge: **how can we learn versatile and informative data representations that can be effectively reused across diverse downstream applications?**

To address this issue, **self-supervised learning (SSL) has emerged as a key paradigm** [Gui+24]. SSL involves training models—typically neural networks—on *unlabeled* data by solving *pretext tasks*, such as denoising corrupted inputs or distinguishing between perturbed and clean samples. These approaches have achieved remarkable performance, in some cases even surpassing fully supervised methods [Li+25]. However, the mechanisms underlying such success remain only partially understood. In particular, **why** do SSL representations generalize so well, and **what** are the key principles driving this behavior?

A first step toward understanding SSL's performance lies in examining its **connections with dimensionality reduction and optimal transport (OT)**. SSL indeed exhibits strong conceptual and mathematical ties with classical dimensionality reduction techniques such as PCA [AW10], t-SNE [VH08], and UMAP [MHM18], especially in the use of related loss functions and optimization strategies [Dam+23]. On the other hand, dimensionality reduction itself is deeply connected to optimal transport theory—indeed, it can even be interpreted as a special case of OT.

Recent studies [Van+25] have thus demonstrated that SSL and OT share several key principles, suggesting a unifying theoretical framework. OT now plays a central role in a wide range of modern machine learning applications, from generative modeling and domain adaptation to cellular dynamics analysis, neural networks, and graph-based models (see [PC+19] for a detailed overview). Understanding SSL through the lens of OT could thus provide valuable theoretical insights and lead to more principled and efficient algorithms for representation learning.

Intership topic

Based on these recent connections, the objectives of this internship are twofold — theoretical and practical. From a theoretical perspective, the work will build upon the recent research conducted by the supervisors to **understand the theoretical foundations of modern SSL methods through the lens of OT and to develop new, efficient SSL techniques inspired by these findings**.

The practical component of the internship will focus on implementing and applying these algorithms to the analysis of single-cell data. Since the late 2010s, major technological advances in molecular and cellular biology have given rise to single-cell biology, a field enabling genome-wide analysis of molecular data, such as DNA, RNA, and proteins—at the resolution of individual cells. Single-cell datasets typically comprise large multivariate distributions, with thousands to millions of cell and thousands of molecular features. Finally, the internship will also provide an opportunity to contribute to the Python library `POT (Python Optimal Transport)`.

Practical informations

We are looking for a highly motivated student, willing to continue with a PhD thesis, with a background in mathematics (optimization, probability and statistics) and/or electrical engineering (signal/image processing, harmonic analysis). Strong abilities in computer sciences will be appreciated. Experience with Python is also desirable.

The intern will be granted the usual stipend of ≈ 600 euros/month. If the candidate is successful, this internship could be pursued by a PhD **(fundings are guaranteed)**.

The intership is expected to start in **spring 2026** and will be hosted at **Inria Rennes**, within the COMPACT research team. The internship will be co-supervised by **Titouan Vayer (Inria)** and **Franck Picard (CNRS, ENS Lyon)**. Do not hesitate to contact us for more information.

References

[Gui+24] Jie Gui et al. "A survey on self-supervised learning: Algorithms, applications, and future trends". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[Li+25] Jingyang Li et al. "Towards Understanding Why FixMatch Generalizes Better Than Supervised Learning". In: *ICLR*. 2025.

[AW10] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* (2010).

[VH08] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *JMLR* (2008).

[MHM18] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[Dam+23] Sebastian Damrich et al. "From t-SNE to UMAP with contrastive learning". In: *ICLR*. 2023.

[Van+25] Hugues Van Assel et al. "Distributional Reduction: Unifying Dimensionality Reduction and Clustering with Gromov-Wasserstein Projection". In: *Transactions on Machine Learning Research* (2025).

[PC+19] Gabriel Peyré, Marco Cuturi, et al. "Computational optimal transport: With applications to data science". In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.