To be sent by e-mail to myriam.tami@centralesupelec.fr and vincent.mousseau@centralesupelec.fr before the **September 25th, 2023.** The projects' subjects' presentation to students is scheduled for **September 29th, 2023 at 8:15 am.** You will be invited to present your subject quickly and answer any questions they may have. The assignment of projects to pairs of master's students will take place after this presentation.

Subject title	**Benchmarking graph self-supervised learning models and applications**
Contact details of the person(s) proposing the project	**First name: Fragkiskos** **Last name: MALLIAROS** **Mail: fragkiskos.malliaros@centralesupelec.fr** **Web: http://fragkiskos.me/**
Institution (Company, lab,…)	*CentraleSupélec, Centre for Visual Computing (CVN), Inria Saclay (OPIS team)* *https://cvn.centralesupelec.fr/*

Project description

- **Context (1/2 page):**

Recent research efforts in the field of representation learning have focused on the extension of deep learning algorithms to data that is generated from non-Euclidean domains, and generally can be represented as graphs that describe complex relationships among entities (e.g., social, biological or communication networks). To this direction, various representation learning approaches have been introduced to deal with prediction and classification tasks on graphs [1, 2]. Characteristic examples include shallow architectures (e.g., DeepWalk, Node2Vec, EFGE) and Graph Neural Networks (GNNs) models such as Graph Convolutional Networks and Graph Attention Networks [3, 4]. Nevertheless, GNN models primarily focus on (semi-)supervised learning tasks that requires access to annotated (labeled) data. To address this issue, recent research efforts in graph self-supervised learning (SSL) have focused on generating graph representations independently of annotated data. Among different approaches, graph contrastive learning has gain significant attention due to the comparable performance to supervised approaches in various graph representation learning tasks [7, 8].

Typically, in a graph contrastive learning (GCL) framework, several graph views are created through a stochastic augmentation process of the input graph. Then, graph (or node) representations are learned by contrasting positive samples against negative ones. A schematic representation of GCL is given in Fig. 1.

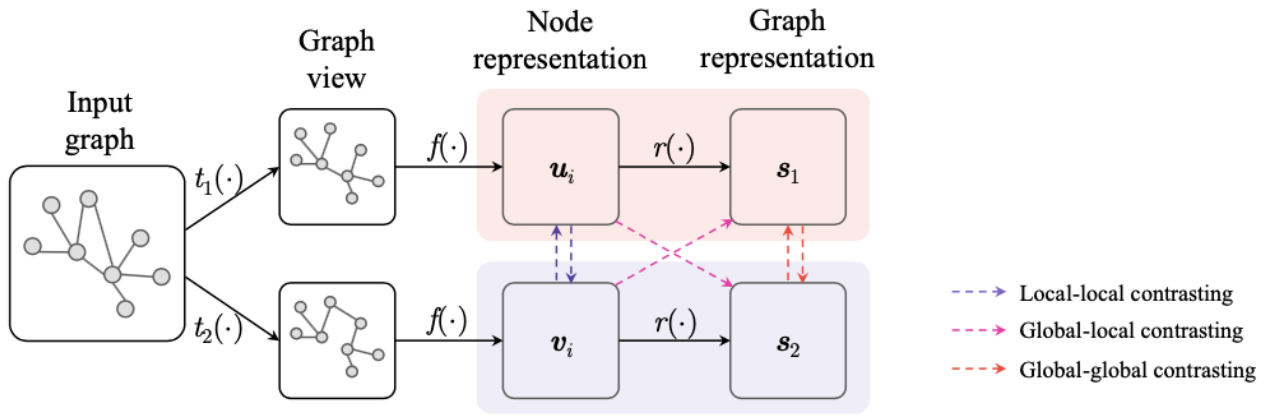

Figure 1: A graph contrastive learning framework [8].

In this project, we aim to benchmark different graph contrastive learning algorithms on biomedical networks targeting a range of different prediction tasks.

- **Goals (1/2 page):**

The objectives of the project are the following:

1. Build a pipeline that allows to compare different graph contrastive learning algorithms. The analysis will focus on (1) different data augmentation strategies, including both topology and feature augmentation; (2) contrastive loss functions (e.g., Information Noise Contrastive Estimation, Jensen-Shannon Divergence, Barlow Twins, Bootstrap Latents); (3) Negative mining strategies.
2. Benchmark the selected models on a wide range of biomedical prediction tasks [5, 6]. One such tasks concerns the problem of *protein function prediction* — a biological task that can be expressed as a node classification task in a protein-protein interaction network. The different steps of the pipeline are shown in Fig. 2.

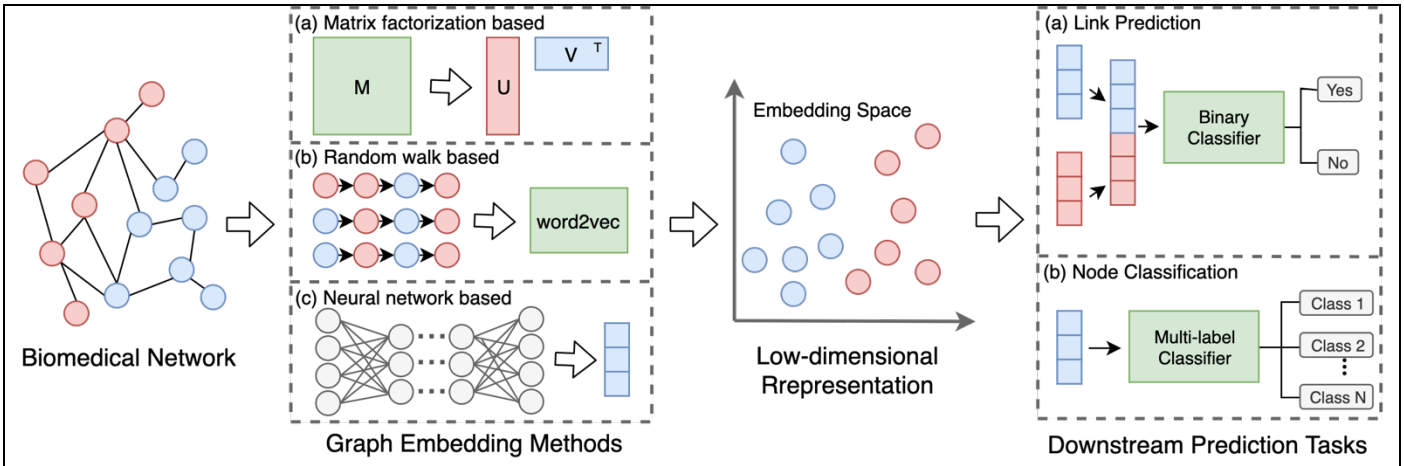

Figure 2: Example of pipeline for graph machine learning to biomedical prediction tasks [6].

We consider that the results of the project will help us understand the functioning of different self-supervised learning algorithms on biomedical networks.

- **Expected work and deliverables**

  - ✓ Familiarity with the related literature.
  - ✓ Identification of graph contrastive models that will be considered in the analysis.
  - ✓ Identification of the datasets and tasks.
  - ✓ Implementation and benchmarking of the models.
  - ✓ (Optional but highly recommended) Publication of the results in a conference/workshop.

The **deliverables** of the project will include (i) a report describing the proposed methodology and experiments; (ii) the source code that allows benchmarking different models (GitHub repository).

- **Technical aspects**

  The students should have a good understanding of machine learning and graph mining techniques, as well as good programming skills (mainly in Python). We expect that the students will reproduce methods/results from the related literature, as well as examine further extensions (in collaboration with us).

Comments (e.g. data access, etc.)
Bibliography [1] Thomas N. Kipf, Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In ICLR, 2017 [2] W.L. Hamilton, R. Ying, and J. Leskovec. Inductive Representation Learning on Large Graphs. In NeurIPS, 2017. [3] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, Yoshua Bengio. Graph Attention Networks. In ICLR, 2018. [4] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. arXiv, 2019. [5] Biomedical datasets: https://github.com/xiangyue9607/BioNEV [6] Yue et al. Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics, 2019. [7] Liu et al. Graph Self-Supervised Learning: A Survey. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2022. [8] Zhu et al. An Empirical Study of Graph Contrastive Learning. In NeurIPS 2021 [9] https://github.com/PyGCL/PyGCL