# Considering symmetries in clustering methods: Application to the study of biomolecules

**Supervisors:**

Cathy MAUGIS-RABUSSEAU
IMT-INSA, Toulouse
cathy.maugis@insa-toulouse.fr

Juan CORTES
LAAS-CNRS, Toulouse
juan.cortes@laas.fr

**Context:**

This internship is part of the project DEFIANT: An interdisciplinary approach to the design of effective nanoparticle-based antimicrobials. It is a collaboration between four laboratories in Toulouse: LAAS, CEMES, TBI and IMT. The goal of this project is to master the design of functionalized gold nanoparticles to improve the therapeutic efficiency of antimicrobials. The starting point for the methodological work is a recent algorithm called IGLOO (Iterative Global exploration and LOcal Optimization), which is developed for the structural prediction of biomolecules on metal surfaces [1, 2]. IGLOO iteratively performs three types of operations: sampling, clustering and local optimization. The DEFIANT project aims at improving and generalizing these three stages to more complex molecular systems.

**Objectives:**

The internship focuses on the question of the development of **unsupervised learning** methods to cluster molecular conformations. The crystallographic structure of the surface induces symmetries on molecular conformations, which can be seen as invariances in molecular positions/orientations. Thus, the challenge consists of proposing a robust unsupervised learning method taking into account the data characteristics for a correct identification of configuration clusters.
The main stages of this internship are:

- Review the state of the art on clustering methods for molecule conformations
- Investigate *ad-hoc* approaches to adapt basic clustering methods such as hierarchical clustering, kmeans-like methods, … to take into account symmetries known *a priori.*
- Investigate general approaches to automatically detect symmetries and to consider them for clustering using more sophisticated methods.
- Evaluate and compare the performances of the different methods on several datasets.

The methods developed during this internship will be implemented by preference in Python.
The clustering methods resulting from this internship will be integrated later to improve IGLOO but also potentially distributed in open-source for the structural biology community.

**Expected skills:**

Strong background in statistics is mandatory, as well as good programing skills (Python).
Background in structural biology is not necessary, but it would be a plus.

**Internship conditions:**

The student will be provided with a monthly stipend of around 550 euros during up to six months.
The student will have an office at the Institute of Mathematics of Toulouseb (IMT), on the campus of INSA Toulouse.

**Applications:**

Please send an email containing your CV to
Cathy Maugis-Rabusseau (cathy.maugis@insa-toulouse.fr) and Juan Cortés (juan.cortes@laas.fr), indicating in the subject “Candidate clustering internship”.

**References:**

[1] S. Abb, N. Tarrat, J. Cortés, B. Andriyevsky, L. Harnau, J. C. Schön, S. Rauschenbach, K. Kern, Carbohydrate Self Assembly at Surfaces: STM Imaging of Sucrose Conformation and Ordering on Cu (100). *Angewandte Chemie*, 131, 8424, 2019.
[2] S. Abb, N. Tarrat, J. Cortés, et al., RSC Advances, 9, 35813–35819, 2019.