Technical assignment for LAMAlab candidates

The purpose of this assignment is **not** that you produce the best possible model.

If you manage to do so, that's great, but it's not the goal and not necessary to pass the assignment.

The purpose is rather to give us a problem we can discuss together. That is, even a partial solution or an attempt to solve the problem is enough–we will discuss it together during the interview.

What we are looking for is your ability to think about the problem, to understand it, to formulate it, and to propose a solution. Under no circumstance you should feel the need to spend more than five hours on this assignment.

The problem

You are given a dataset of around 1000 point clouds in the form of 3D coordinates.

Each point cloud is labeled with an energy value.

Your task is to build a model that predicts the energy value of a point cloud.

You might think of the point clouds as representing the positions of atoms in a molecule, and the energy value as the energy of the molecule. To keep it simple, we only consider one atom type.

Caveats

- Your model should predict the same energy value for a point cloud after performing the following transformations:
  - Rotation: If you rotate the point cloud, the energy value should not change.
  - Translation: If you translate the point cloud, the energy value should not change.
  - Permutation: If you permute the points in the point cloud, the energy value should not change.

Note that there are **many** possible ways to ensure this. You are free to choose any method you like or come up with your own.

- There are a different number of points in each point cloud. Your model should be able to handle this.

The dataset

The dataset is provided in the form of a single file `data.json`. It contains the following variables:

- `X`: A numpy array of shape `(M, n, 3)`. Each row represents a point cloud. Each point cloud consists of n points, each point is represented by 3 coordinates. M is the number of point clouds.
- `y`: A numpy array of shape `(M,)`. Each row represents the energy value of the corresponding point cloud in `X`.

You can load the dataset with the following code:

```python
import json
import numpy as np

with open("data.json", "r") as f:
    data = json.load(f)

y = np.array(data["y"])
X = data["X"]

assert len(X) == len(y)
assert isinstance(X,  list)
```

You can download the dataset from a link that we will provide you with.

Deliverables

We do not expect you to write a report or a presentation. Instead, we would like to receive your code 3 days before the interview. You can send it to us by email or upload it to a private repository on GitHub or GitLab. If you do not provide us with code, please sketch your ideas in a few sentences.

We will then review your code and discuss it together during the interview.