Prediction of nuclear localization

Supervisor: Maria Littmann
Students: Tba
Prediction of subnuclear localization

- Following the idea of LocNuclei, students will predict
  1. Whether a protein is localized only to the nucleus or at least one other compartment
     - Non-Traveler: Proteins only localized to the nucleus
     - Traveler: Proteins also localized to other compartments than the nucleus
  2. Sub-nuclear localization of proteins
     - 13 different compartments
     - Proteins can be localized to more than one compartment

https://www.genome.gov/genetics-glossary/Nucleus
Details

- **Task 1: Traveler vs Non-Traveler Prediction:**
  - Dataset: 1,098 sequences (559 non-traveler, 539 traveler)
  - Output: Binary prediction for each protein whether it is a traveler (1) or not (0)

- **Task 2: Sub-nuclear Localization Prediction:**
  - Dataset: 1,934 sequences
  - Output: Prediction for each of the 13 compartments whether a protein is localized to it (1) or not (0) → proteins can be assigned to more than one compartment (multi-label classification)

- **Input feature:**
  - Both task will be based on the same input feature
  - We will be using **SeqVec embeddings**
    - Representation of a protein sequence as a vector of length 1024
    - Embeddings have been pre-trained using concepts from Natural Language Processing (NLP)
**Literature**


- **Nuclear import and export:**

- **Methods (examples):**