Home assignment - Bioinformatician position

Problem definition

Given a Protein-Protein Interaction (PPI) dataset, create at least two different models that, given an input of two different protein amino acid sequences, will predict whether they will interact or not.

Model's input and output:
Input: protein amino acids sequence A, protein amino acids sequence B
Output: Binding / No binding

Dataset details

The PPI dataset attached is in the format of a Fasta file. Each sample in the file is a pair of two proteins that interact with each other. The IDs of the two are written in the description of the sample, and their sequences are written one after the other, separated by a "-", as the sample's sequence.

Example:

```
>511145.b1424 511145.b1375
MDRRRFIKGSMAMAAVCGTSGIASLFSQAAFAADSDIADGQTQRFDFSILQSMAHDLAQTAWRGA
PRPLPDTLATMTPQAYNSIQYDAEKSLWHNVENRQLDAQFFHMGMGFRRRVRMFSVDPATHLA
REIHFRPELFKYNDAGVDTKQLEGQSDLGFAGFRVFKAPELARRDVVSFLGASYFRAVDDTYQYGL
SARGLAIDTYTDSKEEFPDFTAFWFDTVKPGATTFTVYALLDSASITGAYKFTIHCEKSQVIMDVE
NHLYARKDIKQLGIAPMTSMFSCGTNERRMCDTIHPQIHDSDRLSMWRGNGEWICRPLNNPQK
LQFNAYTDNNPKGFGLLQLDRDFSHYQDIMGWYNKRPSLWVEPRNKWGKGTIGLMEIPTTGET
LDNIVCFWQPEKAVKAGDEFAFQYRLYWSAQPPVHCPLARVMATRTGMGGFSEGWAPGEHYPE
KWARRFAVDFVGGDLKAAAPKGIEPVITLSSGEAKQIEILYIEPIDGYRIQFDWYPTSDSTDPVDMR
MYLRCQGDAISETWLYQYFPPAPDKRQYVDDRVMS-MKSKDTLKWFPAQLPEVRIILGDAVVEVA
KQGRPINTRTLLDYIEGNIKKTSWLDNKELLQTAISVLKDNQNLNGKM
```

Highlighted in blue - the first protein's ID and sequence, and in orange - the second protein's ID and sequence.

Expected results

Deliverabiles:

1. All code files in a dedicated folder
2. The models' files as a pickle
3. Documentation file - in a format of your choice
4. Overview presentation file

All zipped together in one folder.

Please provide:

- An overview presentation meant for presenting the assignment to an audience at Converge, that will include:
  - The models' analysis and results
  - Comparison between the different models.
  - Future steps
- The full code you wrote for this assignment
  - Including explanation and documentation of where and how (which prompts) you used AI-based assistants, if so.

Guidelines

Follow the following guidelines:

- Write the code in Python
- There's no need to use GPU for this, more traditional models will suffice.
- You may use AI-assistance while writing (both code and documentation), but provide the documentation and explanation as stated above.
- Document your research process in addition to the resulting models.