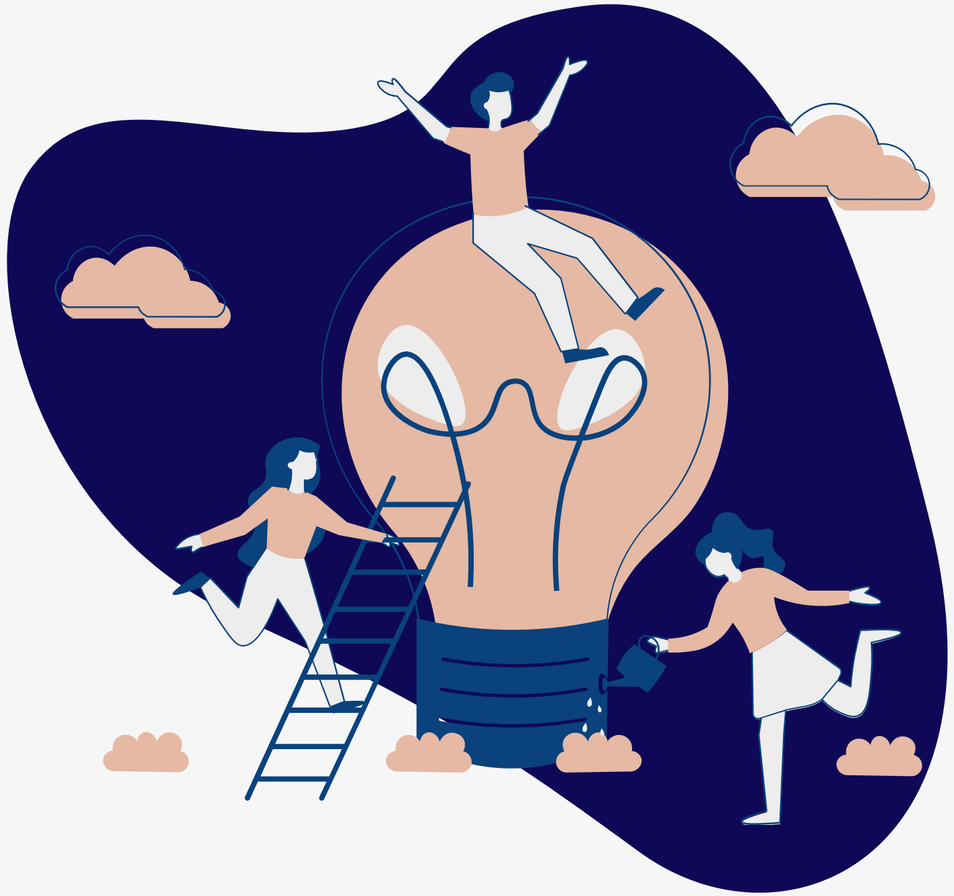
PSCsystem User Study

Three Tasks & Questionnaires

This user experiment aims to validate **whether large language models can enhance the intelligence of multimodal frameworks in protein structure tasks**.

Why are we conducting this experiment?

We aim to understand whether the new PSCsystem Pipeline, empowered by large language models, can offer a more intelligent and efficient solution for multimodal protein-structure-related tasks than two classic Pipelines for



Help us to know you!

Background Survey

As the very first step of the experiment, you will complete a user background survey to help us better understand your background and requirements.

Please enter the user background questionnaire by following the link:
https://forms.gle/3J8UoRuwniXc9RVx7

Experiment Overview

We will conduct three tasks on the aforementioned datasets. Before starting each task, we will provide you with user guides for Baseline, ProteinEngine, and PSCsystem, along with bio examples.

Questionnaire Overview

Background Survey: https://forms.gle/5CpbrrrVtFriZ112A

System Study: Please express your insights into the framework's usability in the SUS questionnaire and provide your feedback on the workload of this experiment in the TLX questionnaire. Please access them through the following links:

- For **Baseline**:
  SUS: https://forms.gle/mpyDWZXhrh3JTnCA6
  TXL: https://forms.gle/PHvk7dcnm5miqrnp6
- For **ProteinEngine Pipeline**:
  SUS: https://forms.gle/nBfCrUDu7nXRDo9cA
  TXL: https://forms.gle/ysfGu6pqxQPMfADy8
- For **PSCsystem Pipeline**:
  SUS: https://forms.gle/zw9KeGfYzR4xHiw98
  TXL: https://forms.gle/jCZTmLkEUbNKsXC96

Task 1

Explore the Baseline Matching Pipeline

Instructions for Use

After downloading the compressed file, please run the Executable File “Baseline.” This will start the code testing. Your command should be in English and Chinese.

1. When you see the following text, it means it’s ready to take your input.

```
Set 1 of 4
Please enter your command (输入您的命令):
```

2. You will be testing one model and one task per time; please enter the model and task names as listed below (‘突变点位预测’ and ‘蛋白逆折叠采样’ are the two task names in Chinese, please copy and paste directly to your command):

Model: ESM1v, ESM-IF, LGN, MSA

Task: inversefolding, mutationhotspot, 突变点位预测, 蛋白逆折叠采样

For example,

```
Set 1 of 4
Please enter your command (输入您的命令): esm1V for INVERSEFOLDING
```

Note: The baseline can only take names that match exactly as elements in model and task lists.

3. Enter a Path and then enter the number, and click enter. You will complete one set of testing.

```
Please enter your command (输入您的命令): ESM1v for inversefolding
Please enter the path to the protein's pdb file: jansn/sjansn.pdb
Please enter the number of samples: 4
Running Model name: ESM1v model for results...
python main.py [Model name: ESM1v, sample, jansn/sjansn.pdb, 4]
*******************************************
Searching summary (搜索总结):
Detected Model: Model name: ESM1v
Detected Task: Task name: inversefolding
Model Match: Yes
Task Match: Yes
Command to execute: python main.py [Model name: ESM1v, sample, jansn/sjansn.pdb, 4]
Total processing time: 13.95 seconds
*******************************************
```

4. Repeat three times. Then, copy the script of testing and submit it to the **SUS baseline matching questionnaire.**

Task 2

Explore ProteinEngine Pipeline

Instructions for Use

After downloading the compressed file, please run **python3 ProteinEngine.pyc** (note that it is a .pyc file) in the terminal. This will start the code testing (please note that in Python, arguments should not contain any punctuation marks such as ' ", ., etc).

During testing, please enter your Python command after "User:", for example:

python3 main.py --* --*** --*****

Each person has a maximum limit of 50 errors. Tests are required to end after all 6 tasks mentioned above are completed, then you can exit!

This is an example of a successful run:

```
Processing
Native sequence loaded from structure file is:
SVLQSLFEHPLYRTVLPDLTEEDTLFNLNAEIRLYPKAASESYPNWLRFHIGINRYELYSRHNPVIAALLRDLLSQKISSVGMKSGGTQLKLIMSFQNYGQALFKPMKQTREQETPPDFFYFSDFERHNAEIAAFHLDRILDFRRVPPVAGRLVNMT
REIRDVTRDKKLWRTFFVSPANNICFYGECSYYCSTEHALCGKPDQIEGSLAAFLPDLALAKRKTWRNPWRRSYHKRKKAEWEVDPDYCDEVKQTPPYDRGTRLLDIMDMTIFDFLMGNMDRHHYETFEKFGNDTFIIHLDNGRGFGKHSHDEMSIL
VPLTQCCRVKRSTYLRLQLLAKEEYKLSSLMEESLLQDRLVPVLIKPHLEALDRRLRLVLKVLSDCVEKDGFSAVVENDLD
Native sequence
Log likelihood: -1.65
Perplexity: 5.20
likehood: -1.6505588966079905
mutant_result[-1.6737752793710443, -1.6570755970628956, -1.6642282992978639, -1.6505588966079905]
```

This is an example of a failed run:

```
You have provided the wrong code, please provide it again
you make an error!
```

After all are successfully run and ended, an output similar to the following will appear. Please record the number of errors (number of input errors) and the running time (in seconds), as well as the total number of attempts.
If everything fails (exceeds the limit), also record the number of errors and the running time. In this case, a total of 50 attempts were made.

Examples for Task 2 (For task specification, please see Workflow.doc)

1. **ESM-IF Model**: Supports two tasks:
   - Mutation Site Prediction Scoring: Requires a protein's PDB file and a corresponding FASTA file.
   - Protein Inverse Folding Sampling Task: Requires a protein's PDB file and the number of samples needed. To execute this task, run main.py with arguments like [esm_if1_gvp4_t16_142M_UR50, sample, User_study_file/protein.pdb, 2].
2. **ESM1v and MSA Models**: Used for mutation site prediction scoring.
   - ESM1v: Requires a protein sequence and a CSV file of the mutations. Run main.py with arguments like [Model, Method, Sequence, Mutant_csv_path].
   - MSA: In addition to ESM1v requirements, it needs an MSA file. Arguments include [Model, Method, Sequence, Mutant_csv_path, Msa_path, Scoring_strategy].
3. **LGN Model**: Used for mutation site prediction scoring. It only needs to specify which protein is being analyzed. The arguments would be [Model, Dataset], where Model is 'lgn' and Dataset is the name of the dataset being used (e.g., GFP).
4. **Gearnet**: For fine-tuning pre-training for downstream tasks. It requires specifying the pre-trained weights to be imported and the dataset on which to fine-tune. The arguments needed are [Model GearNetPretrained_model, GearNetDataset], where Model is 'gearnet', GearNetPretrained_model is the pre-trained weights (e.g., attr), and GearNetDataset is the dataset for testing (e.g., EC).

Task 3

Explore PSCsystem Pipeline

Instructions for Use

After downloading the compressed file, please run the PSCsystem executable. This will start the code testing. The exploration basically focuses on two parts: model and task matching and sample data processing:

1. **Model/Task matching**: Enter natural language command in Chinese or English, and click return to process. The command should include the purpose or the task at least. Irrelevant commands, such as “anything is possible,” will not work for model and task matching:

```
Set 1 of 4
Please enter your command (输入您的命令): anything is possible
Manager Suggestion (AI建议): The phrase "anything is possible" can refer to any of the models and tasks mentioned above,
 as all of them can be used to predict and analyze proteins and protein complexes.
 Processing Time (处理时间): 1.0248980522155762 s
process or not: Y/N
Y
There is no specific model or task mentioned in the given command.

Time for extracting model and task name: 1.4770143032073975s
```

2. Enter natural language commands that contain data or data directories in Chinese or English, and click return to process. If the AI interpretation does not match what you meant, feel free to add clarification or correct your previous command and return.

```
Please enter your data or data path(输入数据或文件路径): Given AA sequence SJJABJSNDIINSNNDBDB
Detected:This input represents direct data (like text) without punctuations. It does not resemble a file or directory pa
th.
Time:0.8496992588043213
Enter clarification:
process or not: Y/N
Y

********Recorded the input*********
"Certainly! Here is part of the result. Token1: 0.425026, Token2: 3.189499, Token3: 2.111793, Token4: 3.782089. For the
complete result, please refer to the generated output file."
Time: 1.3603460788726807s
```

3. After you see ****all task complete****, you can copy everything and submit to the **SUS PSCsystem questionnaire.**

```
****ALL TASKS COMPLETE****
Enter anything to end all sessions
```

Examples for Task 3 Data Input

1. "Use the model to predict the mutant score for a protein, the path to pdb file 'protein.pdb' and fasta file: 'protein.fasta'."
2. "I would like to use the model for an inverse folding task. Here is the path to the pdb file: protein.pdb'. Additionally, sample it twice."
3. "Use the model to predict the mutant score for a protein. Here is the sequence: [sequence], and the mutant csv file is 'result.csv'."
4. "Use the model to perform the task. Here is the sequence: [sequence]. Additionally, the corresponding mutant file is located at 'F7YBW8/F7YBW8.csv', and the msa file path is 'F7YBW8/F7YBW8.a3m'."
5. "Test the EC dataset, using the attr weight to test it."
6. "Use the model to predict mutant score on GFP protein dataset. Here are the pdb file, msa file, and mutant tsv file: 'Dataset/GFP.pdb', 'Dataset/GFP.a3m', 'Dataset/GFP.tsv'"
7. .....

We thank you for your participation and feedback! Your input is valuable to our research!

For further comments or concerns, please contact fh2450@tc.columbia.edu