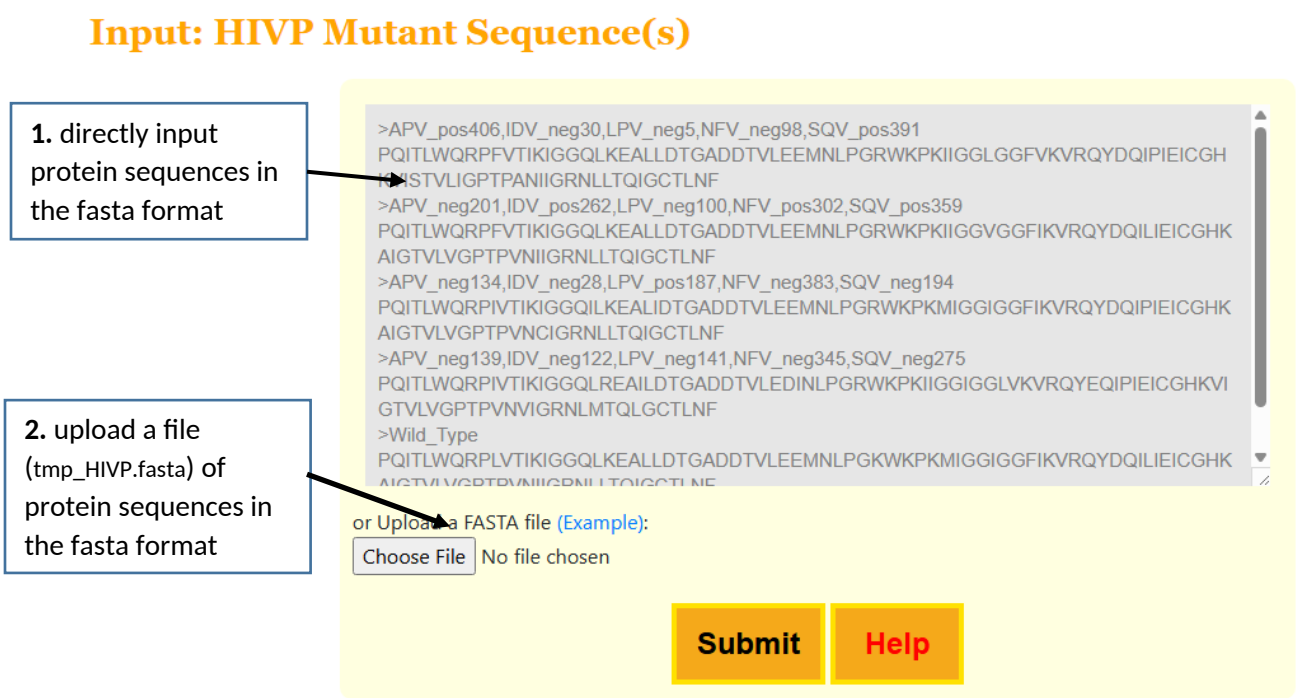
HIVP-LLM Instruction

HIVP-LLM is a protein large language model-based transformer model that can be used for drug susceptibility and resistance prediction in human immunodeficiency virus (HIV) for personalized therapies. HIV-1 is known for its high mutation rate in its protease (HIVP), and the protease susceptibility testing is essential to develop new antiretroviral drugs or optimize the use of existing drugs. HIVP-CGR is built to predict the HIV susceptibility to drug treatment defined by the IC_{50} of a mutant vs. the wild-type control. To this end, we curated a large dataset including 776 unique HIVP mutants against 5 drugs with experimentally determined positive (susceptible to drug treatment) and negative (resistant to drug treatment) testing results. Based on the data, we developed 5 deep learning models to classify HIVP mutants with respect to five protease inhibitors (IDV: Indinavir, SQV: Saquinavir, NFV: Nelfinavir, APV: Amprenavir, LPV: Lopinavir). The model was built based on ProteinBERT to tokenize HIVP sequences, and the web server is implemented with Apache 2.4.58, Python 3.10.20, and Flask 2.0.3.

1. User Input

The input of **HIVP-LLM** is simple and straightforward. There are two ways as shown below:



The users' input of protein sequences must consist of amino acid characters as single uppercase letters in the fasta format. Otherwise, the input will be considered as "illegal"; If

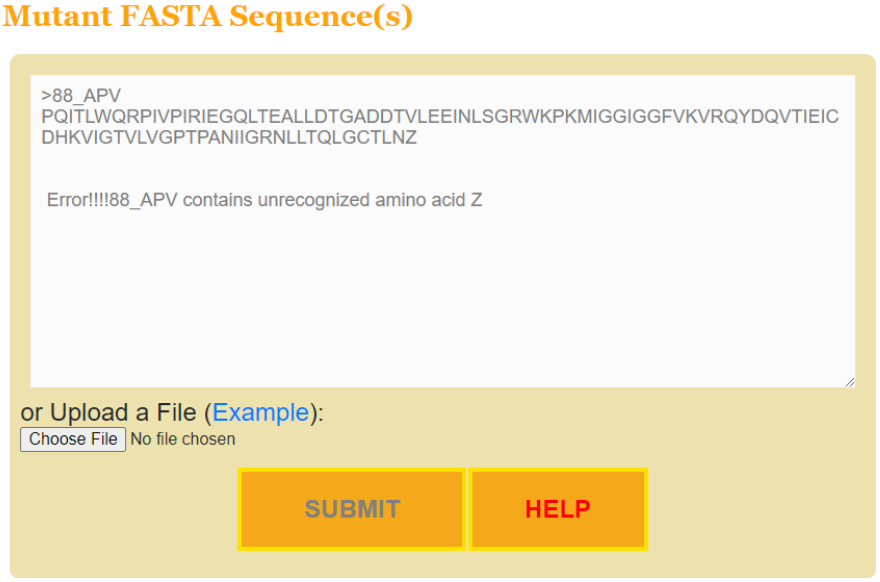
this happens, an error message will show up in the input Form. Here are some examples of incorrect input:

1). Unrecognized amino acid characters

>88_APV

PQITLWQRPIVPIRIEGQLTEALLDTGADDTVLEEINLSGRWKPKMIGGIGGFVKVRQYD
QVTIEICDHKVIGTVLVGPTPANIIGRNLLTQLGCTLNF

An error message will show up to indicate there is an unrecognized amino acid in the sequence ("Z" here).



2). Input without ">" or names for peptides

88_APV

PQITLWQRPIVPIRIEGQLTEALLDTGADDTVLEEINLSGRWKPKMIGGIGGFVKVRQYD
QVTIEICDHKVIGTVLVGPTPANIIGRNLLTQLGCTLNF

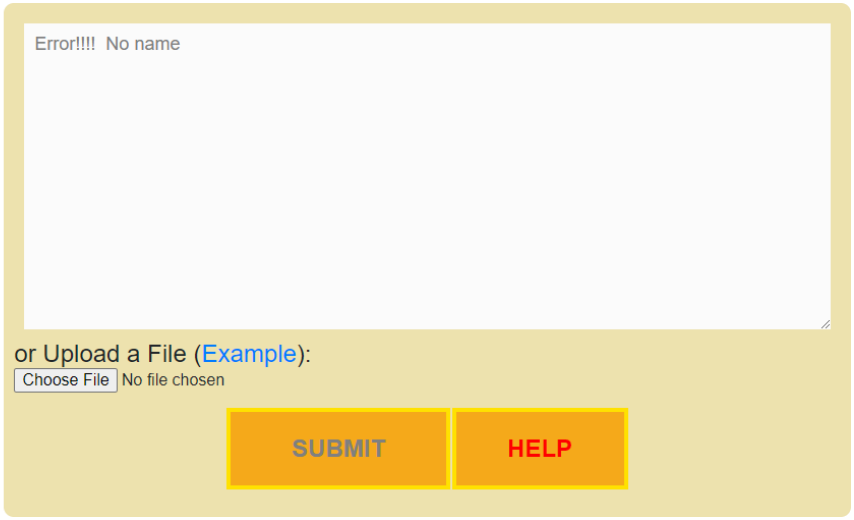
Or

>

PQITLWQRPIVPIRIEGQLTEALLDTGADDTVLEEINLSGRWKPKMIGGIGGFVKVRQYD
QVTIEICDHKVIGTVLVGPTPANIIGRNLLTQLGCTLNF

An error message will show up to indicate there is no name for the sequence.

Input of HIVP Mutant FASTA Sequence(s)

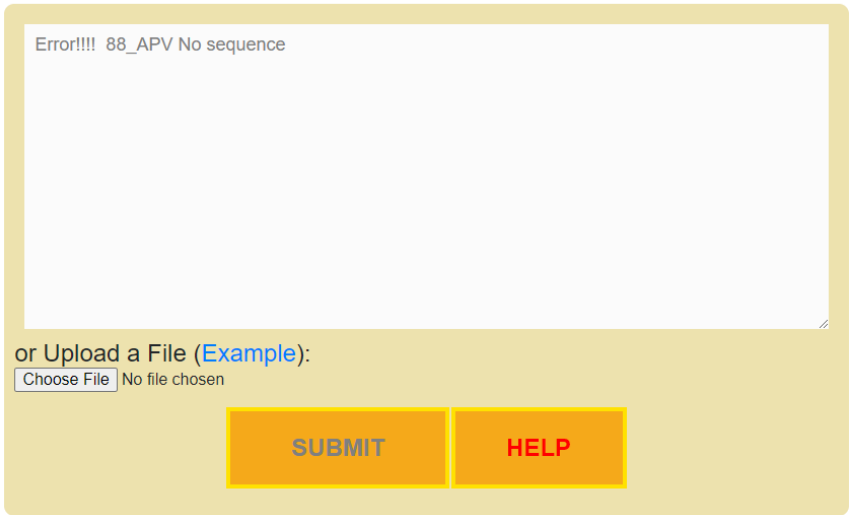


3). Protein name and sequence in the same line

>88_APV QITLWQRPIVPIRIEGQLTEALLDTGADDTVLEEINLSGRWKPKMIGGIGGFV KVRQYDQVTIEICDHKVIGTVLVGPTPANIIGRNLLTQLGCTLNF

An error message will show up to indicate 88_APV has no sequence.

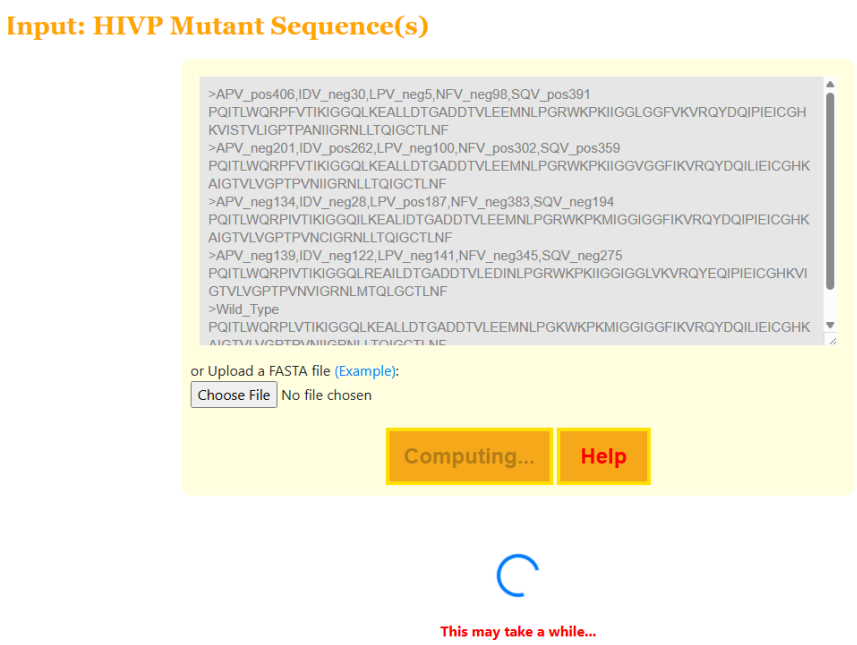
Input of HIVP Mutant FASTA Sequence(s)



It is worth noting the HIVP mutants generally have 99 amino acids in their monomer sequences, except some with insertions or deletions. However, HIVP-LLM can accept any length of protein sequences and make predictions; please be advised that such predictions may not be robust since our models were built mostly based on mutants with 99 residues.

2. Output of HIVP-LLM

Once the job is submitted, The “Submit” button will be changed to “Computing…”, and a blue spinner will show up, together with red “This may take a while…”, as indeed the prediction usually take a while to load the model and parameter and complete the prediction, depending on the number of sequences to predict, as illustrated below.



The output of **HIVP-LLM** is also easy to understand, with six columns in a table. The 1st column is the HIV protein name or ID as provided by users. From the 2nd to the 6th column are the predicted susceptibility of the input sequences with respect to 5 HIVP inhibitors (drugs): APV, IDV, LPV, NFV, and SQV. “1” represents the mutant is susceptible to the drug (the drug us effective to inhibit the virus), and “0” represents the mutant is resistant to the drug (the drug is ineffective to treat the virus). Please refer to our manuscript for more details. The below shows some examples of predictions: the mutant (APV_pos406,IDV_neg30,LPV_neg5,NFV_neg98,SQV_pos391) is susceptible to protease inhibitors APV and SQV, while resistant to the other three drugs (IDV, LPV and NFV). This means if a patient has this HIVP mutant, he/she can be treated with APV or SQV, but not IDV, LPV or NFV. Of note, the wild type HIVP is sensitive to all 5 drugs.Similar interpretations can be applied to other HIVP mutants with respect to the 5 HIVP inhibitors. All such preictions are in agreement with experimental observations.

Output: Predicted Drug Susceptibility (Download Results)

Sequence_ID	APV	IDV	LPV	NFV	SQV
APV_pos406,IDV_neg30,LPV_neg5,NFV_neg98,SQV_pos391	1	0	0	0	1
APV_neg201,IDV_pos262,LPV_neg100,NFV_pos302,SQV_pos359	0	1	0	1	1
APV_neg134,IDV_neg28,LPV_pos187,NFV_neg383,SQV_neg194	0	0	1	0	0
APV_neg139,IDV_neg122,LPV_neg141,NFV_neg345,SQV_neg275	0	0	0	0	0
Wild_Type	1	1	1	1	1

In addition to displaying the result in a table, we also dynamically provide users with an option to **download the prediction data in prediction.csv file**. This is particularly useful if the input number of sequences is large (e.g., >50). The format of the downloaded file is as follow:

```
Sequence_ID,APV,IDV,LPV,NFV,SQV
"APV_pos406,IDV_neg30,LPV_neg5,NFV_neg98,SQV_pos391",1,0,0,0,1
"APV_neg201,IDV_pos262,LPV_neg100,NFV_pos302,SQV_pos359",0,1,0,1,1
"APV_neg134,IDV_neg28,LPV_pos187,NFV_neg383,SQV_neg194",0,0,1,0,0
"APV_neg139,IDV_neg122,LPV_neg141,NFV_neg345,SQV_neg275",0,0,0,0,0
Wild_Type,1,1,1,1,1
```

	A	B	C	D	E	F
1	**Sequence_ID**	**APV**	**IDV**	**LPV**	**NFV**	**SQV**
2	APV_pos406,IDV_neg30,LPV_neg5,NFV_neg98,SQV_pos391	1	0	0	0	1
3	APV_neg201,IDV_pos262,LPV_neg100,NFV_pos302,SQV_pos359	0	1	0	1	1
4	APV_neg134,IDV_neg28,LPV_pos187,NFV_neg383,SQV_neg194	0	0	1	0	0
5	APV_neg139,IDV_neg122,LPV_neg141,NFV_neg345,SQV_neg275	0	0	0	0	0
6	Wild_Type	1	1	1	1	1
7						

3. Instruction and Help

This detailed instruction can be found by clicking on the red-yellow “Help” button next to the “Submit” button.