**DeepSoluE**

**# 1 Description**

Protein solubility is the precondition for its industrial application and functional interpretation. However, the formation of inclusion bodies is still an inevitable roadblock in protein science and industry, where only nearly a quarter of proteins can be successfully expressed in soluble form. Hence, it is imperative to develop novel and highly accurate predictors that enable the prioritization of highly soluble proteins to reduce the cost of actual experimental work.

We developed a novel tool, DeepSoluE, which predicts protein solubility using a long-short-term memory (LSTM) network with hybrid features composed of physicochemical patterns and distributed representation of amino acids. Comparison results showed that the proposed model achieved more accurate and balanced performance than existing tools.

**# 2. Availability**

2.1. Webserver is available at: <http://39.100.246.211:10505/DeepSoluE/>

2.2 Datasets and source code are available at:

<http://lab.malab.cn/~wangchao/softs/DeepSoluE/> and

2.3 Local running

2.3.1 Environment

Before running, please make sure the following packages are installed in Python environment:

gensim==3.4.0

pandas==1.1.3

tensorflow==2.3.0

python==3.7.3

biopython==1.7.8

numpy==1.19.2

For convenience, we strongly recommended users to install the Anaconda Python 3.7.3 (or above) in your local computer.

2.3.2 Additional software requirements

Two additional softwares, namely USEARCH and TMHMM, are needed for DeepSoluE, we did not provide the two tools in the source code packages because of the license restriction. The two methods can be acquired at the following links:

USEARCH: https://www.drive5.com/usearch/

TMHMM: <https://services.healthtech.dtu.dk/cgi-bin/sw_request>

For convenience, please extract the zip file to the **“softs”** folder of the DeepSoluE-master\_source\_code. Of course, you can fix the **tmhmm\_usearch.py** scripts according your environment.

#2.3.3 Running

Changing working dir to DeepSoluE-master\_source\_code, and then running the following command:

**python DeepSoluE.py -i testing.fasta -o prediction\_results.csv**

-i: name of input\_file in fasta format # folder “sequence” is the default file path of the input\_file

-o name of output\_file # folder “results” is the default file path for result save.