JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

autoBioSeqpy: A Deep Learning Tool for the Classification of Biological Sequences

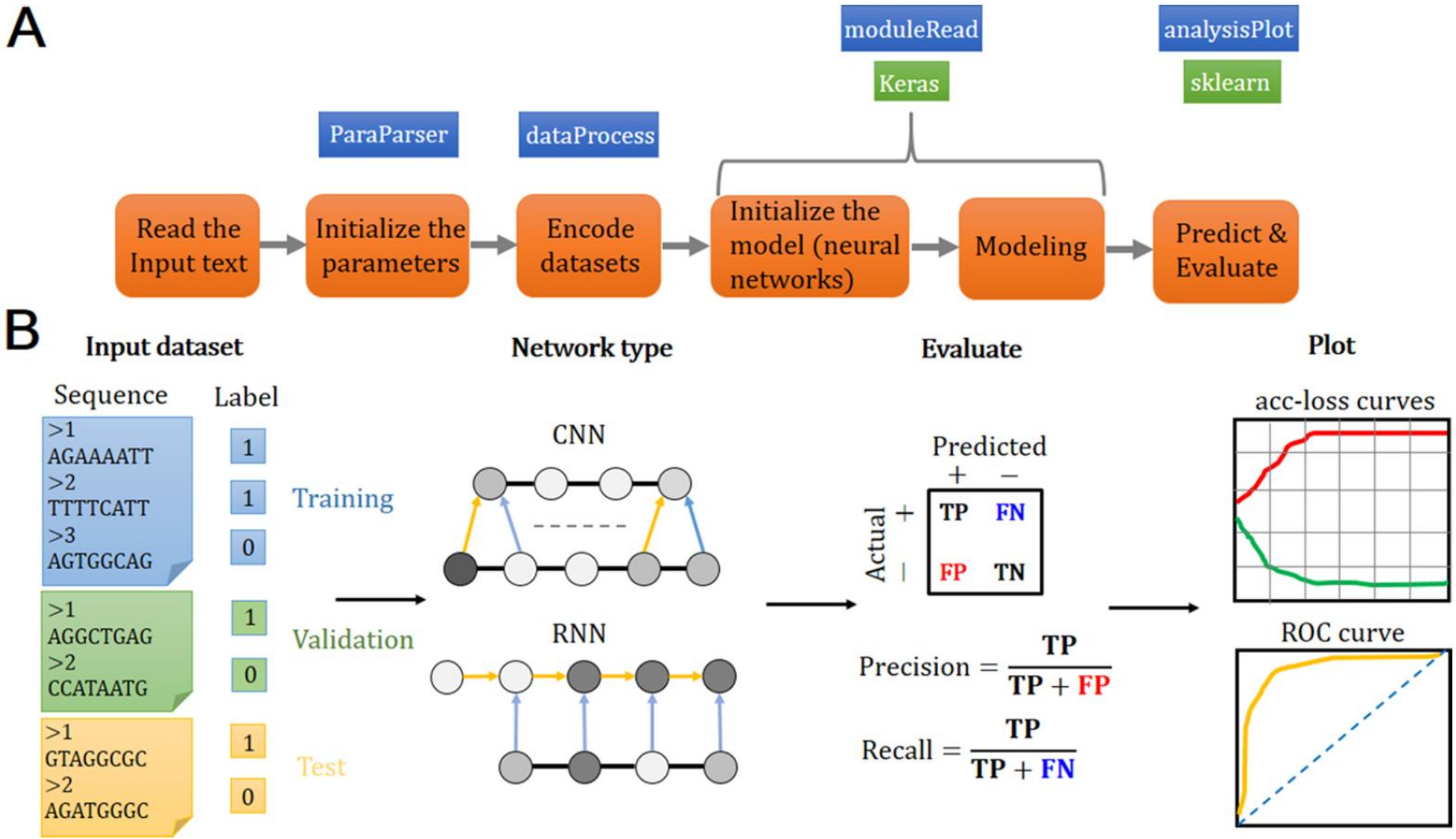
Runyu Jing, Yizhou Li, Li Xue, Fengjuan Liu, Menglong Li,* and Jiesi Luo*

Available tools

- pysster – classification of biological sequences with CNNs
- Selene – implementation of deep learning training/testing on any biological sequence data
- DragoNN – teaching and learning genomics data
- SECLAF – integration of architectures into a webserver for classifying the biological sequences

autoBioSeqpy

- Simplifies the construction and modification of deep learning models
- Users only need to prepare the input datasets
- Data encoding, model development, training and evaluation are run through the command line interface where user may modify the workflow parameters
- Sequence encoding and model configuration are separated into relatively independent parts


A
moduleRead
Keras
analysisPlot
sklearn
ParaParser
dataProcess
Read the Input text
Initialize the parameters
Encode datasets
Initialize the model (neural networks)
Modeling
Predict & Evaluate
B
Input dataset
Network type
Evaluate
Plot
Sequence
Label
>1
AGAAAATT
>2
TTTTCATT
>3
AGTGGCAG
>1
AGGCTGAG
>2
CCATAATG
>1
GTAGGCGC
>2
AGATGGGC
1
1
0
1
0
1
0
Training
Validation
Test
CNN
RNN
Predicted
+
−
Actual
+
−
TP
FN
FP
TN
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
acc-loss curves
ROC curve

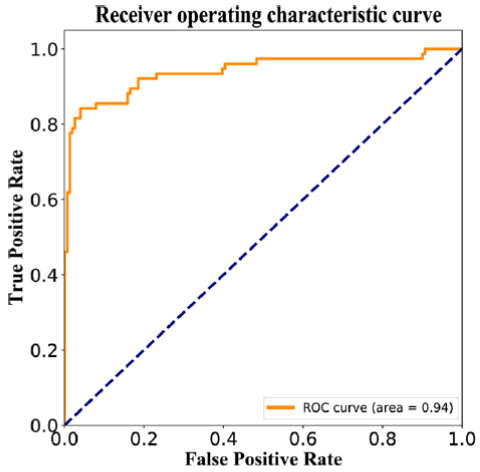
Prediction of Type III Secreted Proteins

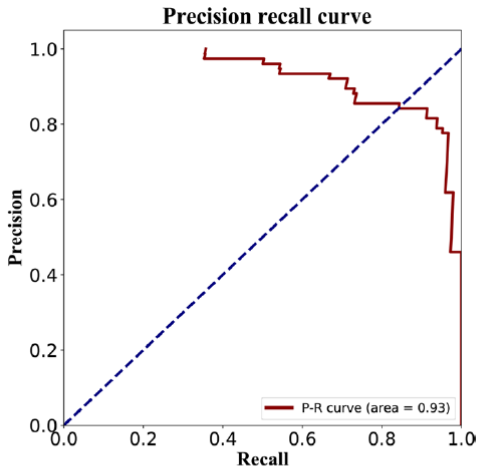
- Dataset: 379 T3SE and 755 non-T3SE
- Only the first 100 residues were used

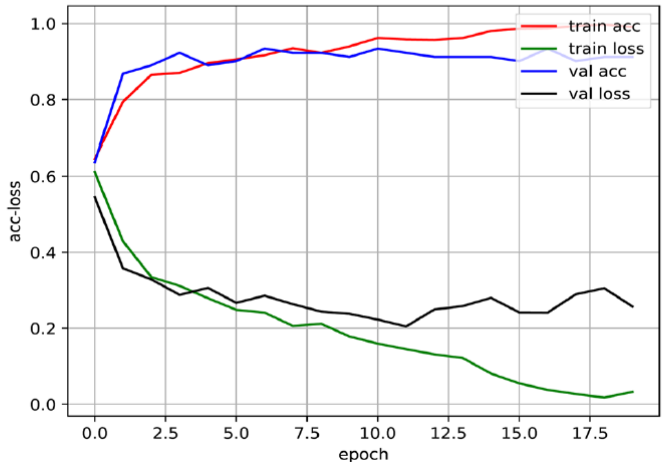
Table 1. Performance Comparison of Different Model Architectures on the 10-Time Test Data set (Case 1 Study)

architecture	encoding	ACC (%)	*F*-value (%)	recall (%)	PRE (%)	*MCC*
CNN-biLSTM		91.8 ± 1.3	86.7 ± 2.6	80.9 ± 6.0	93.9 ± 3.7	0.815 ± 0.030
biLSTM		91.4 ± 0.8	86.4 ± 1.7	81.2 ± 4.9	92.6 ± 3.1	0.807 ± 0.018
CNN	dictionary	90.8 ± 1.6	85.9 ± 2.1	83.3 ± 3.1	89.0 ± 5.1	0.794 ± 0.034
CNN-LSTM		83.1 ± 3.7	70.9 ± 9.2	63.8 ± 13.9	83.1 ± 8.0	0.613 ± 0.088
LSTM		79.6 ± 5.7	61.4 ± 18.4	55.5 ± 24.8	78.3 ± 8.1	0.523 ± 0.156

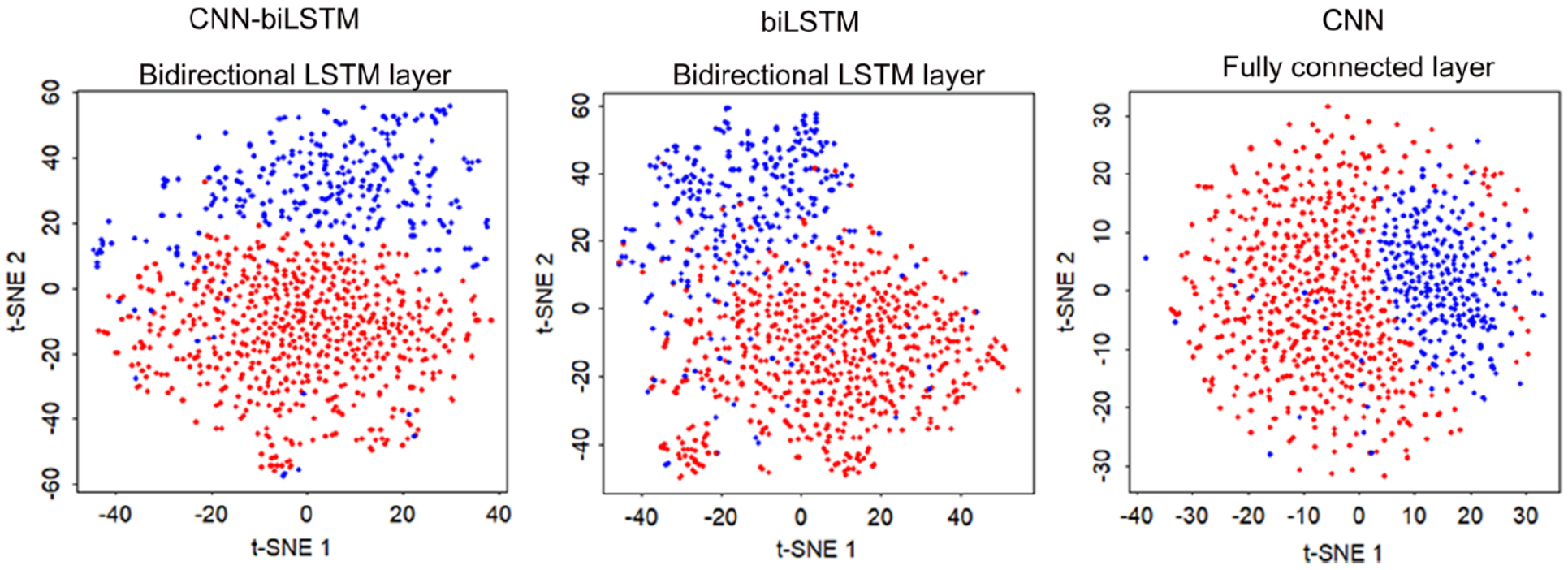
Prediction of Type III Secreted Proteins







Prediction of Type III Secreted Proteins



Prediction of CRISPR/Cas9 sgRNA activity

- Total of 4577 guides were selected to construct three data sets:
  - 2076 guides were for the Wang/Xu data set
  - 1020 guides for the Moreno-Mateos data set
  - 1481 guides for the Doench data set
- Sorting each data set by its cleavage efficiency
- Top 20% - high activity guides and bottom 80% - low activity guides
- Input sequence length: 30

Prediction of CRISPR/Cas9 sgRNA activity

Table 2. Performance Comparison of Different Encoding Ways on the 10-Time Test Data set (Case 2 Study)

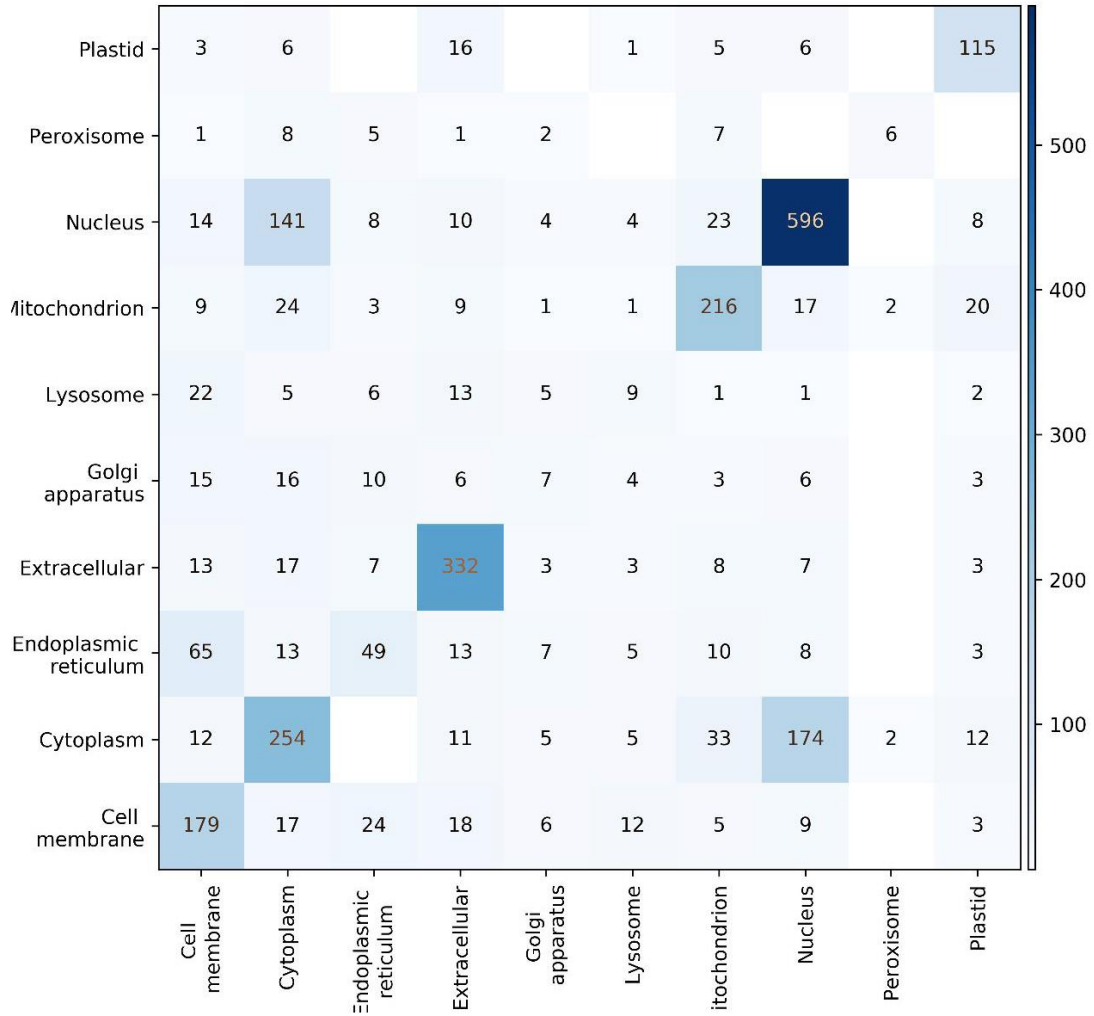
k-mer	encoding	ACC (%)	*F*-value (%)	recall (%)	PRE (%)	*MCC*
Doench data set						
not used	dictionary	82.4 ± 1.9	82.5 ± 2.2	83.1 ± 4.6	82.0 ± 2.6	0.649 ± 0.038
2		78.7 ± 4.7	78.1 ± 5.3	76.5 ± 7.4	80.1 ± 4.6	0.577 ± 0.094
3		78.7 ± 3.2	78.5 ± 3.0	77.4 ± 4.1	79.8 ± 4.8	0.576 ± 0.064
not used	one-hot	80.5 ± 2.1	80.3 ± 2.0	79.5 ± 4.1	81.4 ± 4.2	0.612 ± 0.043
2		79.5 ± 1.8	79.3 ± 2.6	79.2 ± 6.7	79.9 ± 3.2	0.593 ± 0.034
3		79.0 ± 3.7	79.3 ± 3.9	80.8 ± 5.8	78.0 ± 3.4	0.582 ± 0.075
Wang/Xu data set						
not used	dictionary	80.6 ± 3.5	81.2 ± 3.4	83.7 ± 4.8	79.0 ± 4.1	0.615 ± 0.070
2		78.1 ± 2.4	77.7 ± 2.7	76.6 ± 5.6	79.3 ± 3.7	0.566 ± 0.048
3		76.9 ± 3.0	76.4 ± 3.4	74.9 ± 5.3	78.1 ± 3.5	0.539 ± 0.061
not used	one-hot	77.3 ± 3.7	77.1 ± 4.2	76.5 ± 6.7	78.0 ± 4.3	0.550 ± 0.073
2		79.5 ± 4.1	79.3 ± 4.6	78.9 ± 7.1	80.0 ± 4.3	0.593 ± 0.080
3		75.7 ± 3.3	75.8 ± 3.4	76.6 ± 5.8	75.4 ± 3.9	0.516 ± 0.067
Moreno-Mateos data set						
not used	dictionary	72.3 ± 4.2	71.3 ± 6.1	70.2 ± 10.7	73.8 ± 4.7	0.453 ± 0.081
2		75.9 ± 4.3	77.1 ± 4.4	81.7 ± 7.8	73.6 ± 5.4	0.526 ± 0.083
3		69.6 ± 3.3	69.9 ± 4.1	71.2 ± 9.5	69.6 ± 4.4	0.399 ± 0.066
not used	one-hot	73.0 ± 2.8	73.3 ± 3.9	74.9 ± 8.5	72.5 ± 3.2	0.466 ± 0.055
2		74.3 ± 3.6	74.5 ± 4.2	75.9 ± 8.5	74.0 ± 4.3	0.491 ± 0.067
3		74.0 ± 5.5	73.6 ± 4.9	72.4 ± 6.2	75.5 ± 7.0	0.484 ± 0.111

Prediction of protein subcellular localization

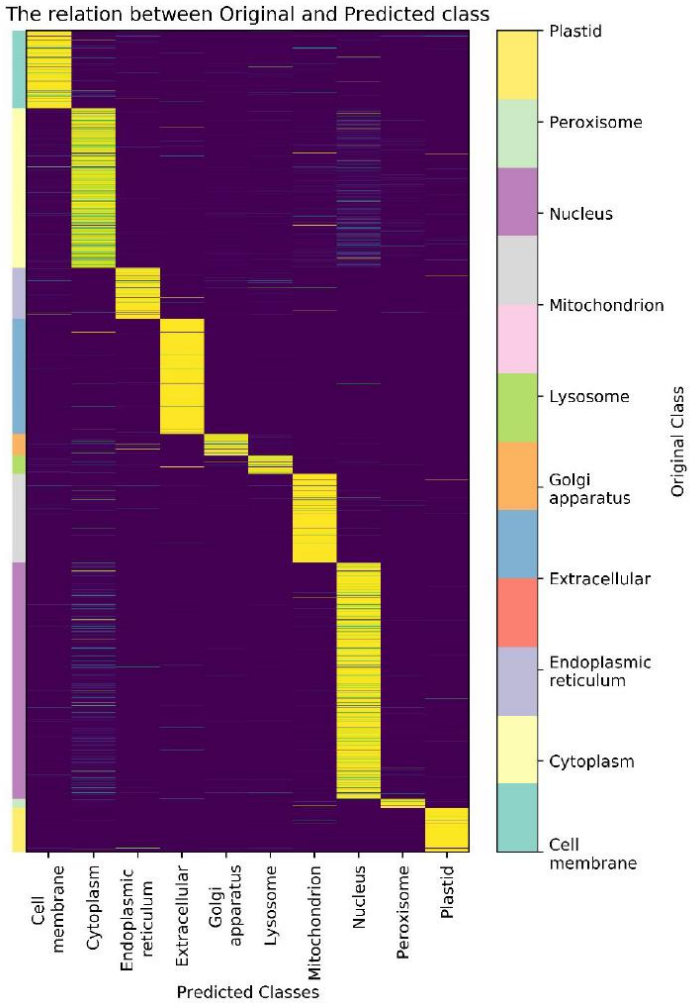
- Dataset obtained from the previous work of Almagro Armenteros et al.
- 11,231 sequences were collected as the training data set and 2773 sequences as the test data set
- divided into 10 classes

Class Name	Training	Test
Cell membrane	1067	273
Cytoplasm	2180	508
Endoplasmic reticulum	689	173
Golgi apparatus	286	70
Lysosome/Vacuole	257	64
Mitochondrion	1208	302
Nucleus	3235	808
Peroxisome	124	30
Plastid	605	152
Extracellular	1580	393

Prediction of protein subcellular localization



Prediction of protein subcellular localization



Summary

- autoBioSeqpy is an easy-to-use tool for designing, training and evaluating deep neural networks for classification of biological sequences
- Authors provide use cases and examples to help users understand deep learning models and the use of this tool