Supplementary Information

Predicting the conformations of the silk protein through deep learning

Supplementary Note 1. Graphical Guide for Calculating Conformation of Proteins through Deep Learning

Perquisites

* Python 3 (stable on 3.7.4 / IPython 7.8.0)
* TensorFlow 2 (stable on 2.3.1)
* NumPy (stable on 1.18.4)
* SciPy (stable on 1.4.1)
* MatPlotLib (stable on 3.2.1)
* Scikit-Learn (stable on 0.21.3)

To start the program, run `amid1_main.py` with Python. On Windows, it usually can simply be done by double-clicking on it.

File I/O

The program only reads spectra in CSV file format. The first row of the CSV file must declare the wavenumber points of every single spectrum. For each row, a spectrum whose wavenumber points consist of ones in the first row is recorded. The first column can be an index column or something else, discarded while parsing. Leaving the first column to be the one for the highest wavenumber point would also be fine.

Be sure to set output file names. Press the “Default” button to set a default name.

Usage

Tab Extract Local: Extract a specific range of spectra to make a new CSV file. Must specify start/stop/step values.

Tab Use AI: Apply a trained model to determine phase contents in each spectrum.

* Model: Standard/Light. Standard represents model 6 in the publication, while Light represents model 8, costing less time.
* Autodetect Trash: whether to use another model specially designed for verifying the existence of β-sheet to filter out once without it.

Tab Gaussian Fit: traditional Gaussian deconvolution method to solve phase contents.

* Area Values: whether to include areas of integrals for 3 phases in the output CSV file.
* Autodetect Trash: same as the one in the “Use AI” tab.
* Parallelism: whether to use multiple cores to compute, which will usually be faster.
* Maximum RMSE Allowed: to filter out any spectra whose error in fitting becomes too large. Usually, a value between 0.03 to 0.07 is rational.

Tab Clustering: Do principal component analysis (PCA) and clustering for input data. Each column represents one label, for example, the wavenumber point, to be a dimension in the procedure. Outputs will be saved in a new directory.

* Number of Clusters: how many clusters will be made.
* New Dir: whether to create a directory for storing the outputs.
* Clusters (.csv): whether to create a CSV file including labels of clusters.
* Visual (.png): show the distribution of the samples on the two principal dimensions that contribute most to the analysis.
* Eigen (.png): if analyzing a set of spectra, show the two curves serving as the “eigenvalues” in PCA.

Tab 2D Distribution: Reconstruct 2D image for input data. Form the second column in the CSV file, each column is regarded as a feature of the data, and a figure is generated for it. The data linearly recorded in the CSV file will be reshaped into a rectangular area in each output image.

* Size: length and height of the rectangle. The product of the two integers must equal the number of records in the input.
* Order: Regular or Alphabetical. To reconstruct 4,096 records to 64 by 64 area, for example, selecting “Regular” means to count from 1, 2, 3, ... to 64 in the first row, and 65, 66, ... in the second, and so on; “Alphabetical” means that it is reconstructed by alphabetical order, such as 1, 10, 100, 1000, 1001, 1002, ... in the first row.
* Matrix File (.mat): writing a matrix file for the reconstructed image with each entry to be a value in the corresponding record. This format works well with certain applications such as MATLAB.
* Image Files (.png): producing images show 2D distributions of each feature.

Showcase

We included “showcase.csv”, a set of spectra for showing the usage. The outputs generated by it were moved to the directory “showcase_output”. Here is how they are produced:

* Click on “Use AI”, click “Browse” to select “showcase.csv” in the line of “Input

File", click "Default" to set a default name for the output, leave all options default, and finally click "Predict by AI" and wait until the program ends.

* Click on "Gaussian Fit". As the input file is the same as what has been dealt with, no changes need to be made for it. Click "Default" in the line of "Output File" to set a default output name. Click "Gaussian Fit" and wait for the termination of the program.
* Click on "Clustering", check if the input file is still the same one, set a default output directory name, and then select Do Clustering/PCA Analysis.
* Click on "2D Distribution", set the input as the file produced in "Use AI" and "Gaussian Fit" steps respectively in two different runs, set default output directory name, unselect "Matrix File", and click "Reconstruct 2D Distributions".

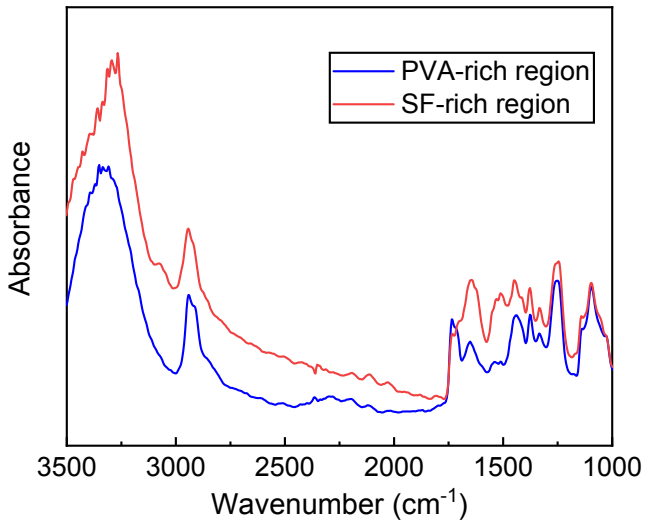

Fig. S1 The single-pixel FTIR spectra extracted from the PVA-rich region and SF-rich region.