**Supplementary file 1**

It is a supplementary description of the main text methods. Many of which are implemented in the specific technical details. In addition to the specific implementation of these methods by using code in R or python language, please refer to (https://github.com/yujuan-zhang/ProtLoc-mexl).

**Amino acid feature construction and Domain knowledge**

**1 Feature representation and calculation method of amino acid composition of protein sequences**

Protein sequence amino acid composition features were calculated for each protein amino acid sequence with 20 amino acid frequencies (AAC), polar amino acids (Polar), non-polar amino acids (Non\_Polar), acidic amino acids (Acid), and basic amino acids (Basic).

/

: The frequency of each amino acid of the protein, for example， indicates the frequency of glycine in this protein.

: Total number of each amino acid of the protein.

: The total number of all amino acids in the protein.

The 20 amino acid composition of proteins is the basis of sequence composition and different amino acids have different physicochemical properties. It has been shown that there is a correlation between amino acid composition of proteins and subcellular localization [1]. In addition， features such as amino acid frequencies are also commonly used to predict protein subcellular localization [2, 3].

**2 Feature representation and computational methods for Stoichioomics features**

Stoichioomics features was performed by calculating the frequency of chemical elements (Oxygen, Carbon, Hydrogen, Sulfur, Nitrogen) of side chain motifs on each protein amino acid sequence separately [4, 5]. The calculation equation is as follows.

: The frequency of each elemental contents of the protein, for example indicates the frequency of carbon elements of this protein.

: The number of atoms of each element of the amino acid side chain on the one protein sequence, for example indicates the number of Carbon on the amino acid side chain.

: The total number of amino acids in the protein.

Chemical elements such as Oxygen, Carbon, Hydrogen, Sulfur, and Nitrogen are the basis for the composition of proteins and play important functions in cellular metabolism, growth, and development. Chemical elemental stoichiometry studies the relationship between the elemental composition of biological macromolecules including nucleic acids and amino acids with different side chain motifs and the evolution of cellular adaptation to the environment ，and proposes the study of environmental limitation of resources [6, 7]. Specifically, the basis of cellular adaptation to the environment requires proteins to perform their corresponding functions in the corresponding subcellular localization compartments. Moreover, based on previous studies in our laboratory, we found that the elemental content of amino acid side chains based on protein sequence is different between different subcellularly localized proteins[4]. Considering the relationship between elemental content and protein function, it is also important to study the elemental content and composition of different localized proteins.

**3. Physicochemical feature representation and calculation methods**

By using the module of the package(version 1.79), we calculated a total of 13 physicochemical Features including molecular weight, aromaticity, isoelectric point, secondary\_structure\_fraction\_α, secondary\_structure\_fraction\_β1, secondary\_structure\_fraction\_β2, molar\_extinction\_coefficient\_disulfid\_bridges, instability\_index, three indicators of protein flexibility(mean, maximum, minimum) and gravy. The specific calculation of each feature can be found in the source code of the module of the package (https://biopython.org/). The physical property of each feature and the role it plays in protein localization or structural function are briefly described below. (Note that the implementation and explanation of the computational process for each of the feature below are mostly referenced in package or other papers. Therefore, please refer to the explanation of this package for the exact calculation and explanation）

**3.1 Molecular weight:**

The molecular weight of a protein reflects the size of the protein molecule and related to the number of amino acids it contains.

**3.2 Aromaticity:**

According to Lobry’s 1994 calculation of the aromatic value of the protein, it is only the relative frequency of 。[8]

Where is the relative frequency of amino-acid of kind in the protein and when the amino-acid is aromatic () and otherwise.

For Physical meaning of aromaticity see Burley’s article[9].

**3.3 Isoelectric point:**

Protein are amphidromical ionized in solution. Assuming a protein is present in a solution, When, , the number of positive and negative ions dissociated from the polar group of the protein is equal, and the net charge is 0. At this time, the value of the solution is the value of the protein. The value of a protein is specific, and is related to the protein structure, but not to the environmental . In proteins and peptides, this depends on the dissociation constant () of the seven amino acids at the end of the polypeptide and the charged groups of and groups. For specific calculation method, please refer to the source code of module. The isoelectric point reflects the dissociation level of protein in the electrolytic solution, and is also related to protein localization and function.[10]

**3.4 secondary\_structure\_fraction\_α, secondary\_structure\_fraction\_β1 and β\_Fold (β2) secondary\_structure\_fraction\_β2:**

The original name for these three features.

secondary\_structure\_fraction\_α indicate the Spiral fraction (α Corner).

secondary\_structure\_fraction\_β1 indicate the β Number of corners (β1).

secondary\_structure\_fraction\_β2 indicate the β Number of β\_Fold (β2)

Calculate fraction of helix, turn and sheet.

Returns a list of the fraction of amino acids which tend to be in Helix, Turn or Sheet.

Amino acids in helix（secondary\_structure\_fraction\_a）: include number of V, I, Y, F, W, L. (Note: V etc. include the following capital letters which are abbreviations of the 20 amino acids)

Amino acids in Turn（secondary\_structure\_fraction\_b1）: include number of N, P, G, S.

Amino acids in sheet（secondary\_structure\_fraction\_b2）: include number of E, M, A, L.

Returns a tuple of three floats (Helix, Turn, Sheet).

The physicochemical properties of these three structures see these references[11-14].

**3.5 molar\_extinction\_coefficient\_disulfid\_bridge:**

Calculate the molar extinction coefficient.

Calculates the molar extinction coefficient assuming cysteines (reduced) and cystines residues (Cys-Cys-bond). The specific calculation method can be found in the source code of module and refer to these papers[15, 16].

**3.6 instability index**:

A statistical analysis of 12 unstable and 32 stable proteins by Rogers et al. (1986) showed that there are dipeptides that occur with significantly different frequencies in unstable and stable proteins. So， we chose to look for the occurrence of 400 possible dipeptides in these two types of proteins. Assuming that the components of dipeptides are independent events, the expected (probable) occurrence of dipeptides can be calculated by the following equation:

Where and are the expected and observed incidence of dipeptide xy, respectively; and are the observed incidence of amino acids x and y, respectively, and T is the total number of amino acids in a particular class.

The chi-squared value between the observed and expected incidence of dipeptide (xy) for each class of proteins was calculated by the following equation:

The mean of the cardinal values for each type of protein was calculated using the formula:

Then was used as the confidence limit to select significant dipeptides for each class of proteins, where the significant dipeptide conditions used to distinguish between stable and unstable proteins were:

And

was also calculated from the observed occurrence of dipeptides in unstable and stable classes of proteins, respectively, using the following equation:

where and are the observed incidence of dipeptides in the unstable and stable classes of proteins, respectively. We chose the value of as the third group of dipeptides that satisfy the following conditions:

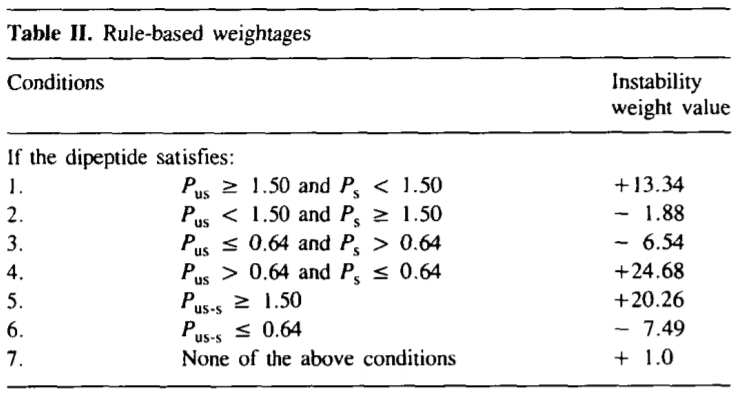
The potential incidence of each dipeptide in all three groups, , was calculated by the following equation:

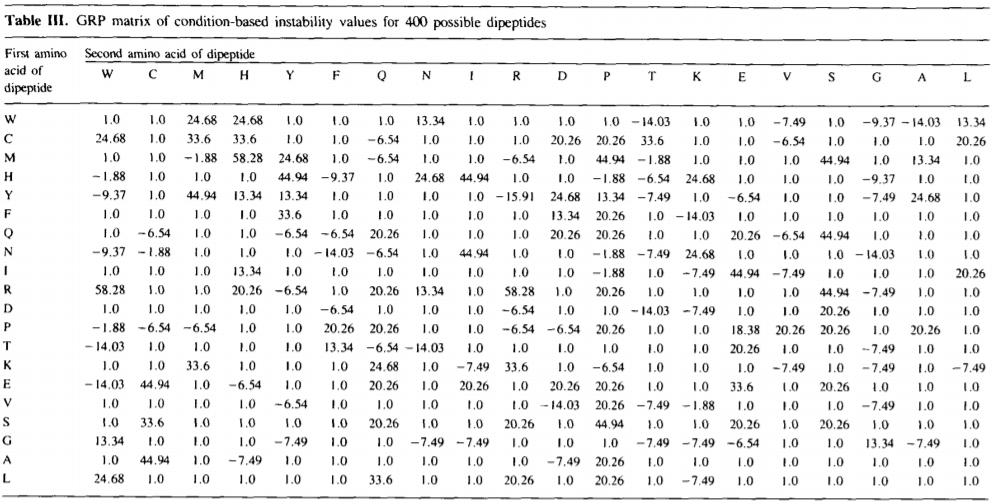
The above three groups of dipeptides were further divided into two subgroups according to the significant difference of from unity (>1.5 or <0.64), and the conditions given in **Table II** were formulated according to the corresponding point positions. The dipeptides satisfying each condition given in **Table II** were classified according to the cardinal values and of each class of dipeptides, respectively. The of all the dipeptides (xy) satisfying the condition were summed to obtain the corresponding instability weight values for each condition. The impact factor for the ith condition was thus estimated by the following equation:

where and are the normalized values of the occurrence of dipeptides satisfying the condition in the unstable and stable classes of proteins, respectively. The influence factor of the condition is operated to bring it into the positive range, which is called the instability weight value, and is given by the following equation:

where LIF is the lowest observed influence factor. The contribution of each dipeptide to instability is obtained by summing the instability weight values corresponding to the conditions satisfied by the dipeptide, called dipeptide instability weight values (DIWV), and the DIWVs of all 400 combinations are represented as matrices in **Table III**, and then the protein instability index (II) is calculated using the DIWV with the following equation:

where is a dipeptide, L is the length of the sequence, and 10 is the scale factor.



**

The difference in the II value () for a pair of tripeptides 'axb' and 'ayb' obtained by changing the central residue and keeping the neighbouring residues the same was estimated by the relation:

We have analysed the data on substitutions in the central position of a tripeptide along with their associated II value, the formula that can be used for estimating the change in the II of a protein would be:

Where is the difference in II of a protein of length L before and after the replacement of residue x residue y.

The instability index is a measure of protein stability, indicating the tendency of a protein to denature or unfold under certain conditions. It is calculated based on the difference in free energy between two tripeptides with different central amino acids. A higher instability index indicates greater protein instability and a greater propensity for denaturation or unfolding. Therefore, the instability index is a useful metric for understanding the impact of amino acid substitutions on protein structural stability and function.

**3.7** **Calculate the protein flexibility indices (mean, minimum, maximum)**

Calculate the protein flexibility indices first, then calculate mean and select minimum, maximum of protein flexibility indices outcome. For protein flexibility indices calculate detail see below（conference from .

Now that window size nine was found to be optimal in predictions with normalized B-values a new equation was determined:

Where .

Higher flexibility indices indicate that a protein is more flexible and dynamic, and may be more likely to participate in specific biological processes, such as protein interactions, structural transitions, and functional regulation.

**3.8 gravy.**

The specific calculation method can be found in the source code of module.

In addition in Informal reference help materials can reference from Jack Kyte and Russell.F.Doolittle’s paper, *A Simple Method for Displaying the Hydropathic Character of a Protein*, which discusses the hydrophobic GRAVY fraction of various types of membrane proteins，which presented in (https://github.com/yujuan-zhang/ProtLoc-mexl).

**GO feature representation and theme clustering**

**1 Doc2vec model.**

In this paper, we use the model in the library of python (version 4.2.0). In this article，only case2, i.e., multiple type of subcellular localization protein is used to conduct feature vector representations for GO annotation information in MF and BP, respectively.

The annotation information of BP and MF are input separately. In addition, all proteins in bench dataset with non-missing BP and MF annotation information excluding the proteins in case2 of the test set, were used to build the corpus (BP and MF were built separately), and using these corpuses to fit pre-training model. Specifically, the model hyperparameters (same for BP and MF) were set as follows: vector\_size=100, min\_count=2, epochs=40, window=5, dm=0 (dm means defining the training algorithm. The default is dm=1, using 'distributed memory' (PV-DM), and dm=0, using distributed bag of words (PV-DBOW). (In addition, PV-DBOW model is similar to skip-gram in word2vec[17, 18]).

Then, the MF and BP annotation information of the training and test sets of case2 are transformed with the above pre-training model to obtain 100 feature vectors for MF and BP, respectively.

**2 GO feature meaning analysis and specific implementation of kmeans** **clustering (cosine distance)**

For detail, Specifically, first the protein GO feature vectors are normalized by l2 parametric, then using package (the R language package with version 2.1.4) to implement based on the Euclidean distance to clustering (This method is equivalent to cosine distance-based clustering). Moreover, before clustering randomly initializing centroid of data and using the average silhouette width method from function to auto choose optimal clusters for clustering.

**3 LDA topic analysis and cluster evaluation**

After obtaining the different protein clusters, the next step is to extract the themes semantics for each cluster, for which we used (Latent Dirichlet Allocation) [19] to model the themes of the protein GO annotation information under each cluster and constructed themes for all clusters in 3 directions and kept the 10 inductive information (keywords) with the largest p-value in each direction. Specifically, we used the package of R language (package version: 0.9.8.1) to perform topic analysis for all clusters separately and set up 3 directional topics, and used the function of the package to obtain the probability of each inductive information (keywords) belonging to different topics. At last, keeping the 10 inductive information (keywords) with the largest P-value in each direction. In addition, the higher overlap or higher probability value of the 3-direction inductive information (keywords) of a topic indicates that the semantic information extracted from this topic is more concentrated.

For measuring the concentration of semantic information extracted by different clusters. In addition to the method of LDA analysis mentioned above, we also calculate the average value of cosine similarity between proteins in the same clusters, then using the similarity matrix of some randomly selected proteins for visualization.

The specific implementation method is achieved by calculating the cosine similarity between two proteins in different clusters using the R language package (package version 0.73.3) and obtaining the similarity matrix for each cluster. Then calculating the average of the cosine similarity of proteins under same clusters based on the similarity matrix. The larger the average value is, the more concentrated the semantic information extracted from the clusters. The calculation formula is.

: is the average of the cosine similarity of proteins under a certain cluster.

: The similarity matrix obtained by calculating the cosine similarity between the proteins of same cluster, where each protein also calculates the similarity with itself and is equal to 1 and the value is on the main diagonal of the matrix. The number of rows of the matrix is equal to the number of proteins in the cluster. So, subtracting is equivalent to subtracting the value of the main diagonal, and means the number of elements remaining after subtracting the elements on the main diagonal from the similarity matrix

: Summing all the values in the similarity matrix .

Please note that the similar matrix is a square matrix, that is, the number of rows and columns are the same.

For the visualization of the similarity matrix of some randomly selected proteins, we also implemented it in R.

Finally, ProtLoc-Mex1 also used the package [20] (package version: 2.24.0) in R language to independently count the relevance of different clusters, which can be used as a complement to our cosine similarity, in other words a highly similarity between both two computational methods indicates a high semantic concentration of that cluster, and vice versa.

**Statistical methods for distribution of features**

The and (standard deviations) of the features under different label categories were calculated by R language and displayed visually by drawing box plots and histograms with the R package. The function in the R language package was used to compare whether there was a statistically significant difference (p<0.05) between the features of different label groups. In addition， the absolute difference between the means of different label categories was calculated as the difference between the means of the features under different label categories, calculated as follows.

: The absolute difference between the mean value of a certain feature of the label and the value of this feature in other label type.

：The mean value of a certain feature of the label

：The mean value of a certain feature of the non- label.

：Number of label categories.

The correlation test (point two column correlation) between the feature values and the different labels can be explained as the correlation between a certain feature (continuous value) and the discrete variable that distinguishes a certain label (1 if this label, 0 if not this label). The formula was calculated as follows.

：Correlation between a certain feature value under a certain label and a discrete variable that distinguishes a certain label.

：Percentage of protein in total protein under a certain label.

：Percentage of other label proteins in total proteins.

：The mean value of a certain feature of a protein under a certain label， the mean value of this feature for proteins of other label. indicates the standard deviation of this feature for all proteins.

In addition， the results of the point two column correlation calculation are the same as the Pearson correlation calculation when the categorical label is a dichotomous discrete variable and takes the value of 1 or 0.

**Prediction model construction and feature filtering**

ProtLoc-Mex1 constructed random forest models by library to predict protein localization, BP clustering cluster prediction, and MF clustering cluster prediction, respectively. The Gini index and Gini gain [21] were calculated by the random forest model as feature weights and filtered features by weight thresholds (weight thresholds were calculated by grid search as hyperparameters). Where the model calculates the feature weights as follows.

: The Gini index of the feature .

: indicates the number of categories.

: indicates the proportion of category in node .

The amount of change in the Gini index before and after branching of the single decision tree is also called the Gini gain as:

is the Gini index score (i.e., weights) for each feature .

and are the Gini indices of the new nodes on the left and the new nodes on the right, respectively.

The importance of each feature in the tree is

Suppose the random forest has trees and the weight of each feature is

After normalization of weights:

**Reference**

1. Cedano, J., et al., *Relation between amino acid composition and cellular location of proteins11Edited by F. E. Cohen.* Journal of Molecular Biology, 1997. **266**(3): p. 594-600.

2. Zhou, H., Y. Yang, and H.-B. Shen, *Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features.* Bioinformatics, 2016. **33**(6): p. 843-853.

3. Shen, Y., et al., *Critical evaluation of web-based prediction tools for human protein subcellular localization.* Brief Bioinform, 2020. **21**(5): p. 1628-1640.

4. Zhang, Y.J., et al., *Subcellular stoichiogenomics reveal cell evolution and electrostatic interaction mechanisms in cytoskeleton.* BMC Genomics, 2018. **19**(1): p. 469.

5. Li, N., J. Lv, and D.K. Niu, *Low contents of carbon and nitrogen in highly abundant proteins: evidence of selection for the economy of atomic composition.* J Mol Evol, 2009. **68**(3): p. 248-55.

6. Elser, J.J., C. Acquisti, and S. Kumar, *Stoichiogenomics: the evolutionary ecology of macromolecular elemental composition.* Trends Ecol Evol, 2011. **26**(1): p. 38-44.

7. Baudouin-Cornu, P., et al., *Molecular evolution of protein atomic composition.* Science, 2001. **293**(5528): p. 297-300.

8. Lobry, J.R. and C. Gautier, *Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes.* Nucleic Acids Res, 1994. **22**(15): p. 3174-80.

9. Burley, S.K. and G.A. Petsko, *Aromatic-aromatic interaction: a mechanism of protein structure stabilization.* Science, 1985. **229**(4708): p. 23-8.

10. Lee, K., et al., *PLPD: reliable protein localization prediction from imbalanced and overlapped datasets.* Nucleic Acids Res, 2006. **34**(17): p. 4655-66.

11. Dobb, M.G., *α-Helix in Fibrous Proteins.* Nature, 1965.

12. Kim, D.I., S.J. Han, and Y.B. Lim, *Unique behaviour of the alpha-helix in bending deformation.* Chem Commun (Camb), 2022. **58**(27): p. 4368-4371.

13. Porter, L.L., et al., *Many dissimilar NusG protein domains switch between alpha-helix and beta-sheet folds.* Nat Commun, 2022. **13**(1): p. 3802.

14. Struthers, M.D., R.P. Cheng, and B. Imperiali, *Design of a monomeric 23-residue polypeptide with defined tertiary structure.* Science, 1996. **271**(5247): p. 342-5.

15. Kendall, E.C. and D.F. Loewen, *The mechanism of oxidation-reduction potential: The oxidation-reduction potential of cysteine and cystine.* Biochem J, 1928. **22**(3): p. 669-82.

16. Spears, R.J., C. McMahon, and V. Chudasama, *Cysteine protecting groups: applications in peptide and protein science.* Chem Soc Rev, 2021. **50**(19): p. 11098-11155.

17. Tomas, M., et al., *Efficient Estimation of Word Representations in Vector Space*. 2013.

18. Le, Q.V. and T. Mikolov. *Distributed Representations of Sentences and Documents*. in *International Conference on Machine Learning*. 2014.

19. Jelodar, H., et al., *Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey.* Multimedia Tools and Applications, 2019. **78**(11): p. 15169-15211.

20. Yu, G., et al., *GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.* Bioinformatics, 2010. **26**(7): p. 976-8.

21. Li, L. and J. Wang, *Research on feature importance evaluation of wireless signal recognition based on decision tree algorithm in cognitive computing.* Cognitive Systems Research, 2018. **52**: p. 882-890.