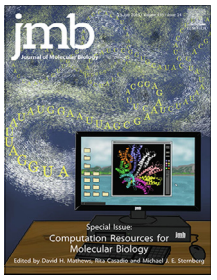

SISPRO: Signature Identification for Spatial Proteomics

Ying Zhou [1†], Yintao Zhang [2,3†], Fengcheng Li [2,3†], Xichen Lian [2,3], Qi Zhu [2], Feng Zhu [2,3*] and Yunqing Qiu [1*]

*1 - **State Key Laboratory for Diagnosis and Treatment of Infectious Disease,** Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Zhejiang Provincial Key Laboratory for Drug Clinical Research and Evaluation, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310000, China*
*2 - **College of Pharmaceutical Sciences,** Zhejiang University, Hangzhou 310058, China*
*3 - **Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University,** Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China*

Correspondence to Feng ZhuYunqing Qiu: The First Affiliated Hospital, School of Medicine, Zhejiang University (YQ. Qiu); College of Pharmaceutical Sciences, Zhejiang University, (F. Zhu). *zhufeng@zju.edu.cn (F. Zhu), qiuyq@zju.edu.cn (Y. Qiu)@zhou36847275 (Y. Zhou), @ZYintao, @zhou36847275 (Y. Zhang)*
https://doi.org/10.1016/j.jmb.2022.167944
Edited by Rita Casadio

Abstract

Spatial proteomics aims for a global description of organelle-specific protein distribution and dynamics, which is essential for understanding the molecular functions and cellular processes in health and disease. However, the application of this technique is seriously restricted by the neglect of robustness among proteomic signatures identified using standard statistical frameworks. Moreover, it is still a major bottleneck to automatically interpretate the identified proteomic signatures due to lack of integration of subcellular information. Herein, an online-tool SISPRO was constructed to (*a*) identify proteomic signatures with good robustness and accuracy via collectively evaluating relative weighted consistency (*CWrel*) & area under the curve (AUC) and (*b*) interpretate the identified signature based on comprehensive subcellular information from 9 organelles and 22 subcellular structures. All in all, SISPRO provides the endeavor to realize the simultaneous improvement of robustness and accuracy in signature identification and the unique capacity in biological annotation lies in its wide coverage of proteins and comprehensive spatial information. SISPRO is expected to be critical in spatial proteomic studies, which can be freely accessed without any login requirement at https://idrblab.org/sispro/

Introduction

Spatial proteomics aims for a global description of organelle-specific protein distribution and dynamics,[1] which is essential for understanding the molecular functions and cellular processes in health and disease.[2] The power of comparative spatial proteomics to reveal the disease mechanisms at the subcellular level has been successfully harnessed by several studies.[3–5] For example, Krahmer *et al.* revealed subcellular reorganization in diet-induced hepatic steatosis using spatial proteomics and phospho-proteomics[6]; Hirst *et al.* reported the role of endosome/lysosome dysfunction in neurodegenerative disorders such as hereditary spastic paraplegia based on subcellular fractionation profiling and quantitative mass spectrometry.[7]

However, the application of spatial proteomics is still seriously restricted by the standard statistical

frameworks, which highly focus on the prediction accuracy of the identified proteomic signatures while neglect the robustness among the signatures from different sub-datasets.[8] An increasing number of studies have recognized that the robustness of feature selection results is equally important as good model performance.[9–11] Moreover, it is still a major bottleneck to automatically annotate the identified proteomic signature due to the lack of integration of subcellular information.[12] It has been widely reported that different pathways and functions are carried out for a certain protein in different subcellular locations.[13,14] The knowledge of the organelle-specific biological annotation has the capacity to help interpret the molecular functions of identified signatures in spatial proteomic studies. Although several popular tools for protein annotation such as Gene Ontology (GO),[15] KEGG[16] and REACTOME[17] greatly contribute to providing the cellular and organism-level functions for proteins, they are not specific to organelles and thus lead to the problem that the annotation from these databases lacks specificity and accuracy for spatial proteomics. In addition, understanding the protein–protein interaction (PPI) network for a certain organelle can also help to discover the molecular basis for human diseases caused by the dysfunction of this organelle.[18] Current PPI data often contain interactions, where the two proteins even don't have the same subcellular localizations. That is to say, the interactions between these proteins may be biophysically possible, but biologically unlikely. Such data can deteriorate the reliability in interactome-based studies.[19] Therefore, it is critical to develop a tool that enables (1) a collective consideration of robustness and accuracy in signature identification[20] and (2) an automated subcellular interpretation for the identified proteomic signatures.[21] No such tool is available yet.

In this study, an online-tool SISPRO (Figure 1) was constructed to (*a*) identify proteomic signature with good robustness and accuracy via collectively evaluating relative weighted consistency (*CWrel*)[22] and area under the curve (AUC)[23] and (*b*) interpretate the identified signature based on a comprehensive set of subcellular annotation information. Particularly, this online-tool first identified the most robust signature list via multiple sampling and then the optimal signature list was determined with highest prediction accuracy and least number of signatures. Second, the organelle-specific biological interpretation for identified signatures were analyzed using the databases constructed in SISPRO that covered a wide range of proteins and comprehensive spatial information (9 organelles and 22 subcellular structures). All in all, these features above make SISPRO distinguished in not only signature identification but also subcellular annotation for spatial proteomics, and therefore SISPRO is expected to have great implications in current spatial proteomic studies, which can be freely accessed without any login requirement at https://idrblab.org/sispro/.

Results and discussion

Workflow and implementation of SISPRO

The workflow of SISPRO included four steps: (I) Data upload and preprocessing. It consisted of missing value imputation, data filtering and normalization. (II) Signature robustness assessment. The *CWrel* was adopted to evaluate the robustness among different signature lists identified via multiple sampling. (III) Assessment of prediction accuracy for identified signatures. Based on the most robust signatures, the signature list with highest AUC and minimal size was discovered for further analysis. (IV) Organelle-specific biological interpretation for identified signatures. For optimal signatures, annotation including organelle- and subcellular structure-based protein function & signaling pathway and protein–protein interaction can be conducted by selecting preferred organelle(s) or subcellular structure(s).

The SISPRO was constructed on a server running *Cent OS Linux v7.4* operating system configured with *Apache HTTP web server v2.4.6* and *Apache Tomcat servlet container*. SISPRO provided a user-friendly interface, which was developed by *R* package shiny v0.13.1. A plenty of *R* packages were used in the background processing. Without any login requirement, SISPRO (https://idrblab.org/sispro/) was accessible by various popular web browsers including Google Chrome, Internet Explorer 10 (or later), Mozilla Firefox and Safari.

Identifying proteomic signature of good robustness and accuracy

To demonstrate the superiority of SISPRO in robustness and accuracy for signature identification, 6 benchmark datasets and 10 popular feature selection methods were adopted in this study. The detailed information on these datasets were provided in **Supplementary Table S1**. For each dataset, the SISPRO and traditional workflow based on 10 popular feature selection methods were used, respectively. The detailed workflow by SISPRO was conducted as follows. First, random selection was conducted ten times with 50% of the samples selected each time, and the feature selection was performed for the ten sub-samples individually resulting in ten lists of the ranked features. Second, the *CWrel* for the Top-N features from the ten lists were calculated, respectively. Third, the signatures from ten lists with highest *CWrel* are aggregated and assessed using AUC. As for traditional workflow, the first and second steps were the same as SISPRO. Then, the mean of the AUCs of Top-N

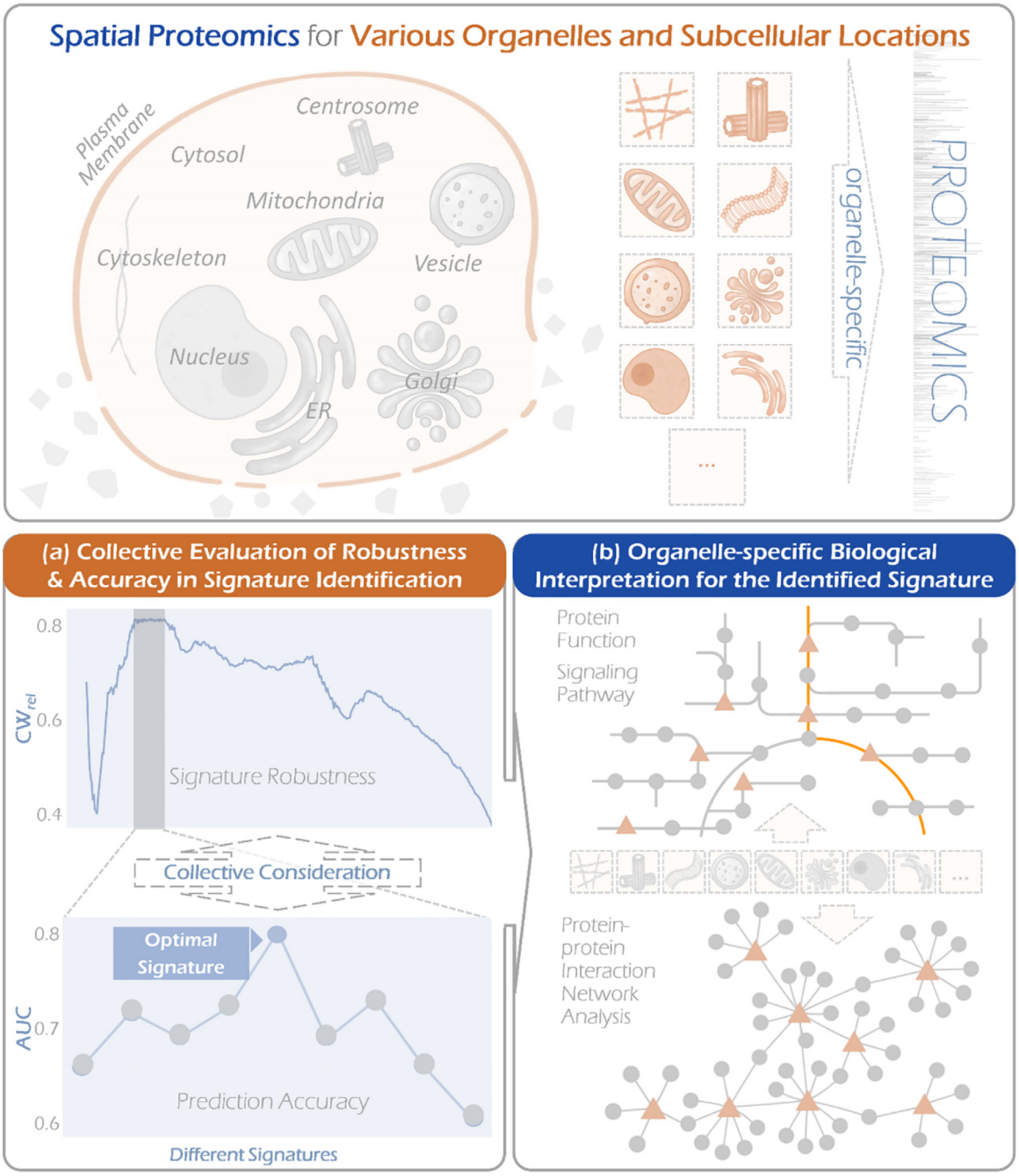

Figure 1. SISPRO was constructed for signature identification and subcellular annotation for spatial proteomics. It was unique in identifying the proteomic signatures with both good robustness & accuracy via collectively evaluating relative weighted consistency (*CWrel*) & area under the curve (AUC) and interpretating the identified signatures based on a comprehensive set of subcellular annotation information (9 organelles and 22 subcellular structures).

features from ten lists were calculated individually, and the robustness of the ten feature lists corresponding to the highest mean AUC was assessed using *CWrel*. The results that compared the accuracy and robustness of the signatures identified by SISPRO and the traditional workflow were described in **Supplementary Table S2**.

As illustrated in Table 1, the AUC and corresponding *CWrel* using SISPRO and traditional workflow for each dataset and feature selection method were provided. In Table 1, the numbers in bold referred to the better AUC or *CWrel* when comparing SISPRO with traditional workflow for a certain feature selection method. Over 85% results from SISPRO were superior than or consistent with that from traditional workflow. For those lower results from SISPRO, the variation between SISPRO and traditional

Table 1 The results of the accuracy and robustness of the signatures identified by SISPRO and the traditional workflow. Six benchmark datasets and 10 popular feature selection methods were adopted in this study. For each dataset, the SISPRO and the traditional workflow for 10 popular feature selection methods were used, respectively. The numbers in bold referred to the better AUC or *CWrel* when comparing the results from SISPRO and traditional workflow for a certain feature selection method.

Data ID	Criterion	AVE (±SD)	Method	Mean	A Variety of Popular Feature Selection Methods Available for Applying									
					CHIS	CBF	ENTROPY	FC	LMEB	PLS-DA	RF	RF-RFE	T-Test	SVM-RFE
JPST000934	AUC	−0.004	SISPRO	0.996	1.000	1.000	1.000	1.000	1.000	1.000	0.987	0.973	1.000	1.000
		(±0.009)	tradition	**1.000**	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**	**1.000**	1.000	1.000
	CW*rel*	+0.250	SISPRO	**0.800**	**0.796**	**0.886**	**0.994**	**0.945**	**0.781**	**0.712**	**0.747**	**0.640**	**0.638**	**0.862**
		(±0.193)	tradition	0.550	0.760	0.844	0.787	0.871	0.373	0.459	0.451	0.022	0.200	0.729
PMID19833877	AUC	+0.142	SISPRO	**0.923**	**0.938**	**0.828**	**0.938**	**0.938**	**0.953**	**0.922**	**0.984**	**0.906**	**0.891**	**0.938**
		(±0.034)	tradition	0.781	0.791	0.766	0.822	0.794	0.788	0.788	0.802	0.767	0.720	0.777
	CW*rel*	+0.131	SISPRO	**0.548**	**0.929**	**0.330**	**0.938**	**0.966**	**0.204**	**0.195**	**0.190**	**0.564**	**0.527**	**0.632**
		(±0.207)	tradition	0.417	0.367	0.278	0.911	0.906	0.195	0.189	0.179	0.093	0.422	0.626
PXD010361	AUC	−0.013	SISPRO	0.977	0.989	0.944	1.000	0.961	1.000	**1.000**	**0.978**	0.944	0.989	0.967
		(±0.026)	tradition	**0.990**	0.989	**0.983**	1.000	**1.000**	1.000	0.967	0.967	**1.000**	**1.000**	**1.000**
	CW*rel*	+0.127	SISPRO	**0.740**	**0.678**	**0.671**	1.000	**0.823**	**0.623**	**0.579**	**0.571**	**0.962**	**0.654**	**0.839**
		(±0.137)	tradition	0.613	0.606	0.404	1.000	0.768	0.476	0.526	0.531	0.484	0.531	0.801
PXD001064	AUC	−0.006	SISPRO	0.718	0.660	0.678	**0.643**	**0.663**	**0.783**	**0.832**	0.657	**0.812**	**0.782**	0.667
		(±0.047)	tradition	**0.724**	**0.705**	**0.708**	0.635	0.651	0.759	0.782	**0.705**	0.767	0.762	**0.762**
	CW*rel*	+0.241	SISPRO	**0.507**	**0.944**	**0.642**	**0.763**	**0.586**	**0.460**	**0.518**	**0.187**	**0.210**	**0.454**	**0.307**
		(±0.234)	tradition	0.266	0.128	0.385	0.331	0.546	0.290	0.270	0.112	0.168	0.310	0.119
PXD003972	AUC	+0.001	SISPRO	**0.983**	1.000	1.000	**0.828**	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(±0.001)	tradition	0.982	1.000	1.000	0.823	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	CW*rel*	+0.205	SISPRO	**0.867**	**0.820**	**0.865**	**0.943**	**0.966**	**0.902**	**0.889**	**0.893**	**0.797**	**0.859**	**0.733**
		(±0.196)	tradition	0.662	0.729	0.422	0.923	0.822	0.747	0.684	0.720	0.142	0.742	0.689
PXD005144	AUC	−0.055	SISPRO	0.929	0.795	0.989	0.803	0.920	**0.999**	**0.999**	**1.000**	0.992	0.996	0.799
		(±0.085)	tradition	**0.984**	**0.984**	0.989	**0.956**	**0.947**	0.996	0.996	0.993	**0.996**	0.996	**0.985**
	CW*rel*	+0.131	SISPRO	**0.619**	**0.794**	**0.732**	**0.581**	**0.697**	**0.629**	**0.629**	**0.317**	**0.532**	**0.716**	**0.559**
		(±0.136)	tradition	0.488	0.463	0.711	0.188	0.659	0.604	0.625	0.230	0.394	0.641	0.362

workflow was very slight. Particularly, the variation within 0.05 and 0.10 were 68.75% and 81.25%, respectively. In addition, the means of AUC and *CWrel* for all feature selection methods based on SISPRO and traditional workflow were also calculated. As shown in Table 1, for all benchmark datasets, the mean robustness of the identified features by SISPRO was superior than traditional workflow with the variation from 0.127 to 0.25. Although the mean accuracy of the identified features for four datasets by SISPRO was lower than tradition workflow, the variation was not significant ranging from 0.006 to 0.055. All in all, these results demonstrated that, compared with traditional workflow, SISPRO has superiority in improving the robustness of identified signatures without sacrificing the accuracy. Such advantages can be of great help to improve the reliability of the identified signature in current proteomics studies.

Biological interpretation using subcellular functional annotation

As shown in the left panel of Figure 2, users can choose their preferred organelle(s) to interpret the subcellular location based biological functions for those identified signatures. A total of 9 organelles and 22 subcellular structures were provided in SISPRO, which greatly satisfied the current needs for spatial proteomics annotation. The annotation result in Figure 2 was conducted based on the signatures from JPST000934 dataset, which were identified using the default pre-processing and feature selection method in SISPRO. The list of proteins with highest AUC and minimal signature size were identified as the optimal signature list. In this case study, two organelles including *Mitochondria* and *Nucleus* were selected for biological interpretation. As shown in Figure 2, different protein functions and signaling pathways that these proteins played in *Mitochondria* and *Nucleus* were identified. The diversity of such results was consistent with previous researches that the same protein could play different functions in different organelles. The biological pathway identified for *Mitochondria* in this study was *oxidative phosphorylation* (OXPHOS) together with the detailed protein function of *complex V component*. The original study of the JPST000934 dataset also reported that restoring OXPHOS had potential in treating hematological malignancy and overcoming chemoresistance,[24] which further validated the accuracy of the identified signatures and organelle-based annotation in SISPRO.

To further demonstrate the accuracy of organelle and subcellular structure-specific biological annotation in SISPRO, the biological relevance of all interpretations for the signatures identified by SISPRO for JPST000934 dataset in different organelles was provided in **Supplementary Table S2,** which was collected from a comprehensive literature review in PubMed. Apart from the OXPHOS and *complex V component* in *Mitochondria*, there were other protein functions and signaling pathways provided for these proteins in *Mitochondria*, such as fusion and fission of ATP5B, mtRNA stability and decay, translation and mitophagy of LRPPRC, iron-sulfur cluster biosynthesis of HSPA9 and so on. Moreover, there were also other biological interpretations in different organelles identified in SISPRO such as angiostatin and MHC class I protein binding function of both ATP5A1 and ATP5B protein in *Cell Membrane*, and lipid metabolism pathway of LRPPRC protein in *Peroxisome*. All in all, these data provide strong evidence of the accuracy and diversity of the biological annotations in SISPRO, which is essential for understanding the molecular functions and cellular processes of proteins in different organelles and revealing the disease mechanisms based on spatial proteomics.

Biological interpretation by organelle-specific PPI network

As provided in the left panel of **Supplementary Figure S1**, users can choose their preferred organelle(s) and subcellular structure(s) to discover diverse PPI networks for those identified signatures in different organelles, which largely fulfils the needs for organelle-specific PPI network analysis. Particularly, PPI network analysis in SISPRO was conducted between the proteins that located in the selected organelle(s) and subcellular structure(s) and the identified signatures. As shown in **Supplementary Figure S1**, the proteins located in *Mitochondria*, *Endoplasmic reticulum* and *Cytosol* were chosen for further analysis. Four signatures identified from JPST000934 dataset based on the default pre-processing and feature selection method in SISPRO were found to have interactions with the proteins located in above three organelles. As illustrated in **Supplementary Figure S1,** the triangles in orange was on behalf of the identified signatures, while the circles denoted the interacting proteins located in the selected organelles or subcellular structures. The edge colors indicated the various organelles or subcellular structures. For a specific signature, the number of proteins with interaction in different organelles varies considerably. To intuitively present the differences of the PPI network in various organelles, the PPI network was conducted individually in a single organelle for the same signatures. As shown in **Supplementary Figure S2**, only three signatures were found to interact with the proteins in *Mitochondria*, while four signatures were discovered to have interactions with the proteins in *Nucleus*, *Golgi apparatus* and *Endoplasmic reticulum*. The most and least number of interactions were in *Nucleus*

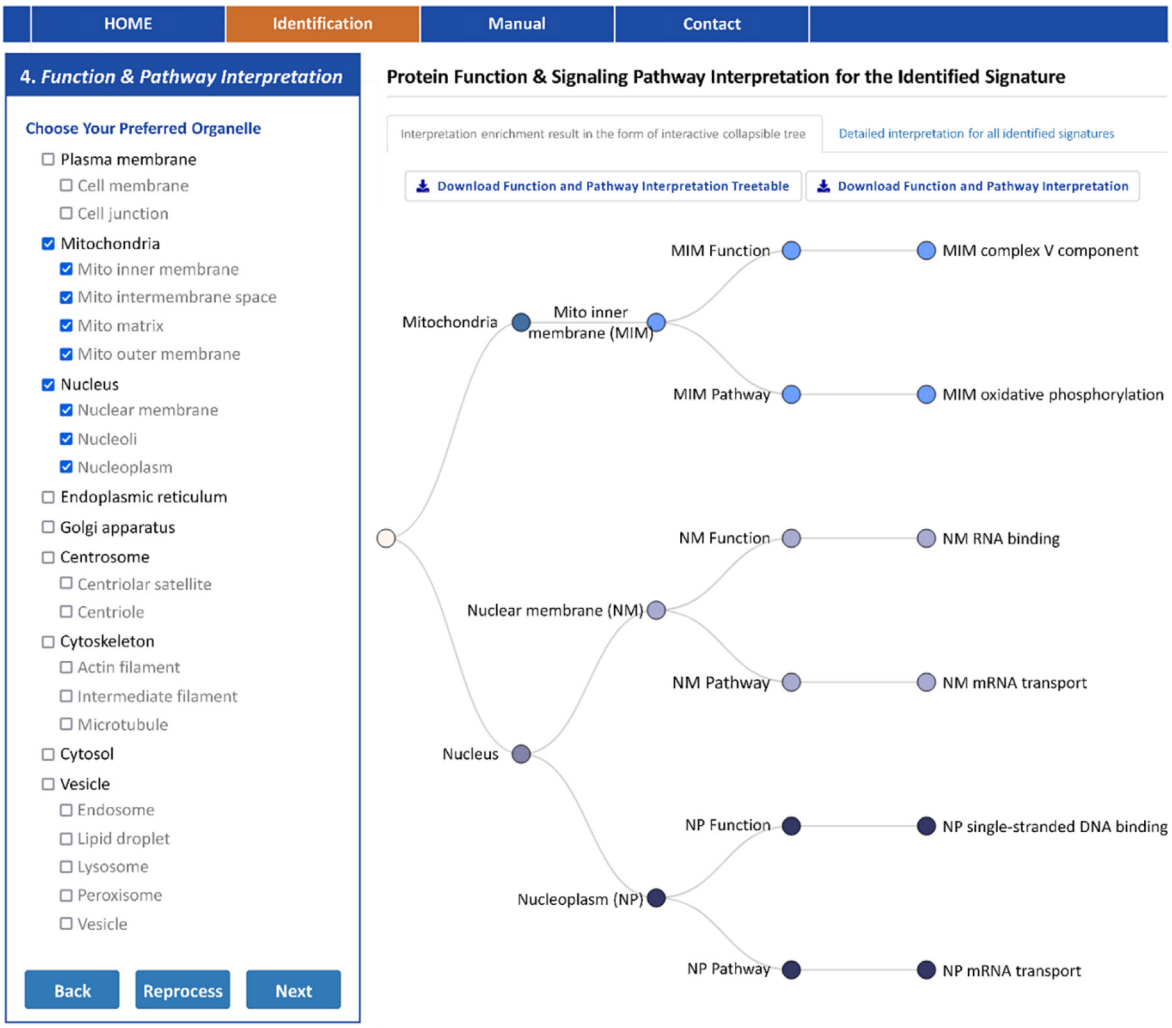

Figure 2. The organelle-specific signaling pathway and protein function interpretation enrichment result. In the left panel, users can choose their preferred organelle(s) to interpret the organelle-specific protein functions and signaling pathways for identified signatures. In the right panel, the enrichment result was presented in the form of interactive collapsible tree, which was categorized by organelles and subcellular structures. All interpretations for these proteins can be found via clicking the 'Detailed interpretation for all identified signatures', which was provided in the form of a table. All results were downloadable online.

and *Endoplasmic reticulum*, respectively. These results above greatly showed the diversity and variation of PPI network in different organelle(s) and subcellular structure(s) conducted in SISPRO and gave us a caution that the localization information of proteins should not be ignored when studying PPI.

Conclusion

In this study, the web server SISPRO constructed for spatial proteomics was unique in both identifying proteomic signature of good robustness and accuracy and interpretating the identified signature based on a comprehensive set of subcellular annotation information. SISPRO provides the endeavor to realize the simultaneous improvement of robustness and accuracy in signature identification and the unique capacity of SISPRO in biological annotation lies in its wide coverage of proteins and comprehensive spatial information. Therefore, SISPRO was expected to be essential and popular in current spatial proteomic studies.

Materials and methods

Collection of benchmark datasets

To assess the performance of SISPRO, six benchmark datasets including 3 spatial proteomics & 3 traditional proteomics datasets were collected to conduct case studies in the **Results and Discussion** section. These datasets were collected from jPOSTrepo,[25] PRIDE[26] and

PubMed.[27] As shown in **Supplementary Table S1**, three spatial proteomic datasets consisted of JPST000934,[24] PMID19833877,[28] and PXD010361.[29] Meanwhile, three traditional proteomic datasets were PXD001064,[30] PXD003972,[31] and PXD005144.[32] Detailed descriptions of these benchmarks were provided in **Supplementary Table S1**.

Performance assessment for signature identification

Robustness evaluating the reproducibility of identified signatures from multiple sampling. Robustness was defined as the reproducibility among multiple lists of signatures identified from different subsets of proteomic data.[20] Recent studies pointed out that the robustness of the identified signatures should be given as much importance as the prediction accuracy,[33] which was regarded as one of the most effective metrics for evaluating the performance of identified signatures.[34] Ignorance of robustness may result in misleading conclusion.[9] In other words, if the lists of identified signatures for a given study were too sensitive to the perturbation in the training data, it would limit the interpretation and practical applications of the results.[33] Thus, the robustness of the identified signatures should be a discriminative criterion for performance assessment.

As more attention has been paid to the robustness of identified signatures, several measures were introduced for robustness evaluation such as *Kuncheva index*,[35] *Dice-Sorensen's index*,[36] *Tanimoto distance*[37] and so on. However, a common problem that these measures met was subset-size-biased.[11] In other words, the larger the selected subset size, the higher values of these measures tended to be yielded.[11] To avoid the subset-size-biased problem, the relative weighted consistency (*CWrel*) was proposed.[38] The *CWrel* represented the overall robustness among multiple lists of signatures, which was calculated based on the occurrence of a specific signature in each set of signatures and the total occurrence of all features in all signature lists.[22] The formula for calculating *CWrel* was as follows:

$$CW_{rel}(S,Y) = \frac{|Y|(N - D + \sum_{f \in Y} F_f(F_f - 1)) - N^2 + D^2}{|Y|(H^2 + n(N - H) - D) - N^2 + D^2}$$

where $|Y|$ represented the total number of all signatures in the original data, S was defined as the set of the n signature lists, f referred to any signature and F_f was the number of occurrences of signature f, N denoted the total number of occurrences of all identified signatures, D was N mod $|Y|$, and H equaled N mod n.

CWrel ranged from 0 and 1, and the *CWrel* value closer to 1 inferred the higher robustness of the identified signatures. Due to its characteristic of independence from the size of the feature subset, *CWrel* was considered as a powerful indicator for evaluating the robustness of identified signatures and therefore adopted in SISPRO.

Accuracy assessing the prediction performance of identified signatures. The goal of feature selection in comparative biological research was to identify a subset of signatures with the potential to correctly discriminate classes of samples (control & case),[39–41] Thus, classification performance of identified signatures was widely used in proteomic studies to demonstrate the reliability of these signatures.[42] The receiver operator characteristic curve (ROC) and area under the curve (AUC) were well-established measures in evaluating classification performance.[43,44] The AUC was between 0 and 1. The higher the AUC value, the better classification performance of the model.[45]

In SISPRO, the AUC was calculated by the following steps. First, the classification model based on the identified signatures was constructed using the support vector machine (SVM) method, where the *R* package *e1071* was used. Specifically, a 5-fold cross-validation was adopted in order to avoid overfitting. The parameters of kernel function, cost, gamma could be defined based on the user's preference. Then, the ROC curve and the AUC value were plotted and calculated via *R* package *pROC*.[46] The list of the signatures with highest AUC and minimum size was selected for next analysis.

Database construction for organelle-specific biological interpretation

Database of subcellular functional annotation. Protein function was tightly associated with its subcellular location.[47,48] For example, IGFBP-2, a major extracellular protein, took part in the insulin growth factor signaling,[49] while its translocation to the nucleus led to vascular endothelial growth factor-mediated angiogenesis.[50] The rapid development of spatial proteomics provided clues for better understanding of protein functions in different locations. Such knowledge has also been applied in multiple fields such as location biomarkers screening,[51] disease mechanism understanding,[52] drug targets identification[53] and drug discovery.[54]

However, inconsistency of protein subcellular location information provided by existing databases and lack of subcellular location specific biological interpretation of proteins limited the further understanding and practical application of signatures discovered in spatial proteomics Thus, a database providing protein subcellular functional annotation was constructed in SISPRO. The construction of the database was mainly divided into two parts: (a) determining of the subcellular location of human proteins; (b) interpretating these proteins with subcellular location specific protein function & signaling pathway information. *First*,

protein subcellular location information from UniProt[55] and Human Protein Atlas[1] was collected and integrated. The different subcellular location entries from the two databases were standardized to 9 organelles and 22 substructures. Due to the variation of the protein subcellular location from the two databases, a comprehensive literature review was conducted to improve the credibility of the data. Specifically, keywords searching including the protein name/gene name and its reported subcellular location from the two databases mentioned above was conducted in PubMed,[27] and publications retrieved for each keywords combination was carefully reviewed. All in all, the subcellular locations of 16,366 proteins in 9 organelles and 22 substructures were identified.

To further provide subcellular location based protein annotation, signaling pathway & protein function information for each protein was first extracted from Gene Ontology (GO) database.[15] Then redundant information was manually removed. To precisely elucidate protein function in specific organelle or substructure, the keywords combination such as "[protein name/gene name] + [GO item] + [subcellular location]", "[protein name/gene name] + [GO ID] + [subcellular location]", "[protein name/gene name] + [subcellular location] + pathway", [protein name/gene name] + [subcellular location] + function" and so on were searched in PubMed.[27] The corresponding results retrieved were carefully reviewed. Finally, 148,116 annotation for 15,562 proteins in 9 organelles and 22 substructures were recorded. Users can choose their preferred organelles and/or substructures to interpret the identified signatures for a given study. The full annotation provided by SISPRO was present in a table and the annotation enrichment result was displayed in the form of a collapsible tree.

Database of organelle-specific protein-protein interaction (PPI). As the basis for most cellular processes, protein–protein interaction (PPI) was regarded as one of the most valuable sources for proteome analysis.[19,56,57] Moreover, increasing evidence has demonstrated that the functions of PPIs were closely related to their spatial distribution and temporal dynamics.[58] For example, interaction between the mitochondrial dynamin-related protein-1 (DRP1) with microtubule-associated TAU protein (MAPT) could trigger excessive mitochondria fragmentation and synaptic defects, leading to neuronal damage.[59]

However, most available PPI data did not take the subcellular location information of interacting proteins into consideration, which would deteriorate the reliability of studies, especially subcellular-specific cellular processes. Thus, a PPI database considering protein subcellular location information was constructed in the following steps. First, the keywords searching including "[protein name/gene name] + interaction", "[protein name/gene name] + PPI", "[protein name/gene name] + protein–protein interaction" was conducted in PubMed.[27] Then the corresponding results were carefully reviewed and the reported PPIs were recorded. The subcellular location information identified in the previous section was also adopted in this section. Based on the PPIs and subcellular location information of proteins, users can find the interacting proteins of the identified signatures that locate in their preferred organelles or substructures.

CRediT authorship contribution statement. **Ying Zhou:** Methodology, Validation, Writing – original draft, Writing – review & editing. **Yintao Zhang:** Methodology, Software. **Fengcheng Li:** Methodology, Software. **Xichen Lian:** Data curation, Visualization. **Qi Zhu:** Writing – review & editing. **Feng Zhu:** Conceptualization, Supervision. **Yunqing Qiu:** Conceptualization, Supervision.

DATA AVAILABILITY

All data in the manuscript are collected and available in PRIDE and PubMed database.

Acknowledgements

Funded by Natural Science Foundation of Zhejiang Province (LR21H300001); National Natural Science Foundation of China (U1909208 & 81872798 & 81971982); Zhejiang Provincial Science and Technology Department Key Technologies R&D Program (2022C03020); Leading Talent of "Ten Thousand Plan" of the National High-Level Talents Special Support Plan of China; Fundamental Research Fund of Central University (2018QNA7023); Key R&D Program of Zhejiang Province (2020C03010); Chinese "Double Top-Class" Universities (181201*194232101); Westlake Laboratory (Westlake Laboratory of Life Science & Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; Information Tech Center of Zhejiang University.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2022.167944.

Received 2 November 2022;
Accepted 27 December 2022;
Available online xxxx

Keywords:
spatial proteomics;
signature identification;
robustness;
accuracy;
organelle-specific biological interpretation

† These authors contributed equally to this work as co-first authors.

Abbreviations:
C*Wrel*, Relative weighted consistency; AUC, Area under the curve; ROC, Receiver operator characteristic curve; SVM, Support vector machine; PPI, Protein-protein interaction; CHIS, Chi-squared test; CBF, Correlation-based feature selection; FC, Fold change; LMEB, Linear models and empirical bayes; PLS-DA, Partial least squares discriminant analysis; RF, Random forest; RF-RFE, Random forest-recursive feature elimination; SVM-RFE, Support vector machine-recursive features elimination; ENTROPY, Entropy-based filters

References

1. Thul, P.J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., et al., (2017). A subcellular map of the human proteome. *Science* **356**, eaal3321.
2. Lundberg, E., Borner, G.H.H., (2019). Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302.
3. Oom, A.L., Stoneham, C.A., Lewinski, M.K., Richards, A., Wozniak, J.M., et al., (2022). Comparative analysis of T-Cell spatial proteomics and the influence of HIV expression. *Mol. Cell. Proteomics* **21**, 100194
4. Bottek, J., Soun, C., Lill, J.K., Dixit, A., Thiebes, S., et al., (2020). Spatial proteomics revealed a CX3CL1-dependent crosstalk between the urothelium and relocated macrophages through IL-6 during an acute bacterial infection in the urinary bladder. *Mucosal Immunol.* **13**, 702–714.
5. Buczak, K., Ori, A., Kirkpatrick, J.M., Holzer, K., Dauch, D., et al., (2018). Spatial tissue proteomics quantifies inter- and intratumor heterogeneity in hepatocellular carcinoma (HCC). *Mol. Cell. Proteomics* **17**, 810–825.
6. Krahmer, N., Najafi, B., Schueder, F., Quagliarini, F., Steger, M., et al., (2018). Organellar proteomics and phospho-proteomics reveal subcellular reorganization in diet-induced hepatic steatosis. *Dev. Cell* **47**, e207.
7. Hirst, J., Itzhak, D.N., Antrobus, R., Borner, G.H.H., Robinson, M.S., (2018). Role of the AP-5 adaptor protein complex in late endosome-to-Golgi retrieval. *PLoS Biol.* **16**, e2004411.
8. Goh, W.W.B., Wong, L., (2018). Dealing with confounders in omics analysis. *Trends Biotechnol.* **36**, 488–498.
9. Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., et al., (2020). Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief. Bioinform.* **21**, 1058–1068.
10. Wang, W., Sue, A.C., Goh, W.W.B., (2017). Feature selection in clinical proteomics: with great power comes great reproducibility. *Drug Discov. Today* **22**, 912–918.
11. Li, F., Zhou, Y., Zhang, Y., Yin, J., Qiu, Y., et al., (2022). POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability. *Brief. Bioinform.* **23**, bbac040.
12. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., et al., (2017). GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11.
13. Meng, D., Yang, Q., Melick, C.H., Park, B.C., Hsieh, T.S., et al., (2021). ArfGAP1 inhibits mTORC1 lysosomal localization and activation. *EMBO J.* **40**, e106412.
14. Ulman, A., Levin, T., Dassa, B., Javitt, A., Kacen, A., et al., (2021). Altered protein abundance and localization inferred from sites of alternative modification by ubiquitin and SUMO. *J. Mol. Biol.* **433**, 167219
15. Gene Ontology C, (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334.
16. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M., (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551.
17. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., et al., (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503.
18. Galletta, B.J., Fagerstrom, C.J., Schoborg, T.A., McLamarrah, T.A., Ryniawec, J.M., et al., (2016). A centrosome interactome provides insight into organelle assembly and reveals a non-duplication role for Plk4. *Nat. Commun.* **7**, 12476.
19. Veres, D.V., Gyurko, D.M., Thaler, B., Szalay, K.Z., Fazekas, D., et al., (2015). ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Res.* **43**, D485–D493.
20. Goh, W.W.B., Wong, L., (2019). Advanced bioinformatics methods for practical applications in proteomics. *Brief. Bioinform.* **20**, 347–355.
21. Rath, S., Sharma, R., Gupta, R., Ast, T., Chan, C., et al., (2021). MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.* **49**, D1541–D1547.
22. Pes, B., Dessi, N., Angioni, M., (2017). Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. *Inform Fusion.* **35**, 132–147.
23. Zheng, X., Xu, K., Zhou, B., Chen, T., Huang, Y., et al., (2020). A circulating extracellular vesicles-based novel screening tool for colorectal cancer revealed by shotgun and data-independent acquisition mass spectrometry. *J. Extracell Vesicles.* **9**, 1750202.
24. Nelson, M.A., McLaughlin, K.L., Hagen, J.T., Coalson, H. S., Schmidt, C., et al., (2021). Intrinsic OXPHOS limitations underlie cellular bioenergetics in leukemia. *Elife* **10**, e63104.

25. Moriya, Y., Kawano, S., Okuda, S., Watanabe, Y., Matsumoto, M., et al., (2019). The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Res.* **47**, D1218–D1224.
26. Perez-Riverol, Y., Bai, J., Bandla, C., Garcia-Seisdedos, D., Hewapathirana, S., et al., (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552.
27. Sayers, E.W., Beck, J., Bolton, E.E., Bourexis, D., Brister, J.R., et al., (2021). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **49**, D10–D17.
28. Hwang, H., Bowen, B.P., Lefort, N., Flynn, C.R., De Filippis, E.A., et al., (2010). Proteomics analysis of human skeletal muscle reveals novel abnormalities in obesity and type 2 diabetes. *Diabetes* **59**, 33–42.
29. Furniss, R.C.D., Low, W.W., Mavridou, D.A.I., Dagley, L.F., Webb, A.I., et al., (2018). Plasma membrane profiling during enterohemorrhagic E. coli infection reveals that the metalloprotease StcE cleaves CD55 from host epithelial surfaces. *J. Biol. Chem.* **293**, 17188–17199.
30. Liu, Y., Buil, A., Collins, B.C., Gillet, L.C., Blum, L.C., et al., (2015). Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11**, 786.
31. Caron, E., Roncagalli, R., Hase, T., Wolski, W.E., Choi, M., et al., (2017). Precise temporal profiling of signaling complexes in primary cells using swath mass spectrometry. *Cell Rep.* **18**, 3219–3226.
32. Saraswat, M., Joenvaara, S., Seppanen, H., Mustonen, H., Haglund, C., et al., (2017). Comparative proteomic profiling of the serum differentiates pancreatic cancer from chronic pancreatitis. *Cancer Med.* **6**, 1738–1751.
33. Li, F., Yin, J., Lu, M., Yang, Q., Zeng, Z., et al., (2022). ConSIG: consistent discovery of molecular signature from OMIC data. *Brief. Bioinform.* **23**, bbac253.
34. Xu, W., Tian, Y., Wang, S., Cui, Y., (2020). Feature selection and classification of noisy proteomics mass spectrometry data based on one-bit perturbed compressed sensing. *Bioinformatics* **36**, 4423–4431.
35. L.I. Kuncheva, A stability index for feature selection, in: Proceedings of the Iasted International Conference on Artificial Intelligence and Applications, 1 (2007) 390–395.
36. Chandrashekar, G., Sahin, F., (2014). A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16–28.
37. Kalousis, A., Prados, J., Hilario, M., (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst.* **12**, 95–116.
38. Somol, P., Novovicova, J., (2010). Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1921–1939.
39. Christin, C., Hoefsloot, H.C., Smilde, A.K., Hoekman, B., Suits, F., et al., (2013). A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol. Cell. Proteomics* **12**, 263–276.
40. Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., et al., (2020). ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief. Bioinform.* **21**, 621–636.
41. Tang, J., Mou, M., Wang, Y., Luo, Y., Zhu, F., (2021). MetaFS: performance assessment of biomarker discovery in metaproteomics. *Brief. Bioinform.* **22**, bbaa105.
42. Conrad, T.O., Genzel, M., Cvetkovic, N., Wulkow, N., Leichtle, A., et al., (2017). Sparse proteomics analysis - a compressed sensing-based approach for feature selection and classification of high-dimensional proteomics mass spectrometry data. *BMC Bioinf.* **18**, 160.
43. McCullough, A.K., Rodriguez, M., Garber, C.E., (2020). Quantifying physical activity in young children using a three-dimensional camera. *Sensors* **20**, 1141.
44. Yang, Q., Li, B., Chen, S., Tang, J., Li, Y., et al., (2021). MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J. Proteomics* **232**, 104023
45. Yang, Q., Wang, Y., Zhang, Y., Li, F., Xia, W., et al., (2020). NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res.* **48**, W436–W448.
46. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., et al., (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* **12**, 77.
47. Orre, L.M., Vesterlund, M., Pan, Y., Arslan, T., Zhu, Y., et al., (2019). SubCellBarCode: proteome-wide mapping of protein localization and relocalization. *Mol. Cell* **73**, 166–182.
48. Narayana Rao, K.B., Pandey, P., Sarkar, R., Ghosh, A., Mansuri, S., et al., (2022). Stress responses elicited by misfolded proteins targeted to mitochondria. *J. Mol. Biol.* **434**, 167618
49. Firth, S.M., Baxter, R.C., (2002). Cellular actions of the insulin-like growth factor binding proteins. *Endocr. Rev.* **23**, 824–854.
50. Azar, W.J., Zivkovic, S., Werther, G.A., Russo, V.C., (2014). IGFBP-2 nuclear translocation is mediated by a functional NLS sequence and is essential for its pro-tumorigenic actions in cancer cells. *Oncogene* **33**, 578–588.
51. Kuechler, E.R., Budzynska, P.M., Bernardini, J.P., Gsponer, J., Mayor, T., (2020). Distinct features of stress granule proteins predict localization in membraneless organelles. *J. Mol. Biol.* **432**, 2349–2368.
52. Seo, J.H., Rivadeneira, D.B., Caino, M.C., Chae, Y.C., Speicher, D.W., et al., (2016). The mitochondrial unfoldase-peptidase complex ClpXP controls bioenergetics stress and metastasis. *PLoS Biol.* **14**, e1002507.
53. Zhou, Y., Zhang, Y., Lian, X., Li, F., Wang, C., et al., (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res.* **50**, D1398–D1407.
54. Mou, M., Pan, Z., Lu, M., Sun, H., Wang, Y., et al., (2022). Application of machine learning in spatial proteomics. *J. Chem. Inf. Model. 62*, 5875-–5895.
55. UniProt C, (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489.
56. Bell, E.W., Schwartz, J.H., Freddolino, P.L., Zhang, Y., (2022). PEPPI: whole-proteome protein-protein interaction prediction through structure and sequence similarity, functional association, and machine learning. *J. Mol. Biol.* **434**, 167530
57. O'Neill, A.C., Uzbas, F., Antognolli, G., Merino, F., Draganova, K., et al., (2022). Spatial centrosome

proteome of human neural cells uncovers disease-relevant heterogeneity. *Science* **376**, eabf9088.

58. Liu, Z., Xing, D., Su, Q.P., Zhu, Y., Zhang, J., et al., (2014). Super-resolution imaging and tracking of protein-protein interactions in sub-diffraction cellular space. *Nat. Commun.* **5**, 4443.

59. Manczak, M., Reddy, P.H., (2012). Abnormal interaction between the mitochondrial fission protein Drp1 and hyperphosphorylated tau in Alzheimer's disease neurons: implications for mitochondrial dysfunction and neuronal damage. *Hum. Mol. Genet.* **21**, 2538–2547.