JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

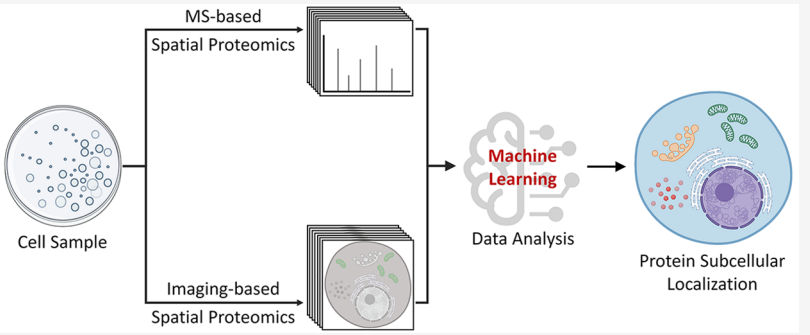
pubs.acs.org/jcim Review

Application of Machine Learning in Spatial Proteomics

Minjie Mou, Ziqi Pan, Mingkun Lu, Huaicheng Sun, Yunxia Wang, Yongchao Luo, and Feng Zhu*

Cite This: J. Chem. Inf. Model. 2022, 62, 5875−5895 Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Spatial proteomics is an interdisciplinary field that investigates the localization and dynamics of proteins, and it has gained extensive attention in recent years, especially the subcellular proteomics. Numerous evidence indicate that the subcellular localization of proteins is associated with various cellular processes and disease progression. Mass spectrometry (MS)-based and imaging-based experimental approaches have been developed to acquire large-scale spatial proteomic data. To allow the reliable analysis of increasingly complex spatial proteomics data, machine learning (ML) methods have been widely used in both MS-based and imaging-based spatial proteomic data analysis pipelines. Here, we comprehensively survey the applications of ML in spatial proteomics from following aspects: (1) data resources for spatial proteome are comprehensively introduced; (2) the roles of different ML algorithms in data analysis pipelines are elaborated; (3) successful applications of spatial proteomics and several analytical tools integrating ML methods are presented; (4) challenges existing in modern ML-based spatial proteomics studies are discussed. This review provides guidelines for researchers seeking to apply ML methods to analyze spatial proteomic data and can facilitate insightful understanding of cell biology as well as the future research in medical and drug discovery communities.

KEYWORDS: *spatial proteomics, machine learning, deep learning, protein subcellular localization, mass spectrometry, imaging, data resources, analytical tools, cell biology*

1. INTRODUCTION

Spatial proteomics is an interdisciplinary field that studies the localizations of proteins and their dynamics.[1] It provides information on spatial protein distribution in organs and tissues at the macroscopic level and subcellular protein mapping at the microscopic level, the latter of which, also termed organelle proteomics or subcellular proteomics, has received preferential attention in recent years.[2,3] The compartments of eukaryotic cells consist mainly of membrane-bound organelles (such as endoplasmic reticulum, Golgi apparatus and mitochondria) and some nonmembranous subcellular structures (such as centrosome and ribosome).[4] The functions of proteins are largely determined by their localizations, especially at the subcellular level, since proteins must be localized to specific compartments and interact with the corresponding components to fulfill their biological functions in cellular processes[5] (for example, apoptosis,[6] signaling[7] and mitosis[8]). Furthermore, it has been shown that protein mislocalization is implicated in various cellular dysfunctions and pathologies, including cancers,[9,10] neurological diseases[11−14] and metabolic disorders.[15,16] The knowledge of spatial proteome is of paramount importance for in-depth understanding of cell biology.

With the rapid growth within this field, a variety of sophisticated experimental techniques have been developed to support large-scale investigations of spatial proteome.[17] Currently, there are three major and complementary spatial proteomic approaches: mass spectrometry (MS) analysis of biochemical fractionated organelles,[18] protein proximity labeling coupled with MS[19] and fluorescent imaging of protein localization.[20] These approaches vary in principle and

Received: September 16, 2022
Published: November 15, 2022

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

experimental design, which makes them suitable for different types of applications. Based on these advanced experimental techniques, a large number of high-quality spatial proteomic data sets have been established. It is essential to effectively manage and deposit these valuable data to ensure data accessibility.[21]

Rigorous data analysis is as critical as data acquisition. To make sense of these spatial proteomic data, elaborate methodologies are needed to analyze the data and get insights into the underlying biological interpretation.[22] Numerous computational methods have been applied to deal with the complexity of spatial proteomic data.[23] Historically, predictions of protein subcellular localization used amino acid sequence and extracted features as inputs to train classifiers.[24,25] It has been reported that the inaccuracy of protein localization inferred from sequence or similarity had a detrimental effect on the identification of protein clusters.[26] Currently, with the development of quantitative MS technology, protein localization profiles based on organelle purification or fractionation have been widely characterized.[27] Statistical analysis has been used to map proteins to specific subcellular compartments, such as Student's *t*-test,[28] Chi-square test,[29] Mann−Whitney test[30] and partial least-squares discriminant analysis.[31,32] With the significant improvement of organelle resolution and proteomic coverage, analysis of the complex data generated by multi-organelle profiling approaches becomes a huge challenge, leading researchers to focus on machine learning (ML) methods.[1] In particular, ML methods (such as unsupervised,[33,34] semi-supervised[35,36] and supervised methods[37,38]) have been widely applied in spatial proteomic data visualization, novelty detection, protein localization prediction and other tasks.[26] Recent substantial advances in microscopy have enabled imaging-based spatial proteomics approaches to simultaneously visualize multiple proteins at subcellular resolution.[39] Subsequently, deep learning (DL) methods are gaining popularity as tools for spatial proteomics image analysis because they can automatically derive features and capture the subcellular image patterns.[40] Furthermore, several advanced modern ML strategies have been applied in spatial proteomic data analysis, such as self-supervised learning,[41,42] transfer learning[43−45] and ensemble strategy.[46,47] Overall, ML has proved successful in spatial proteomics and is playing an increasingly important role in data analysis pipeline.

However, compared with the rapid progress of experimental technologies, the development of ML methods for spatial proteomic data analysis has encountered bottlenecks and is still lagging behind.[48] The sheer diversity of ML algorithms can be daunting for researchers, and no consistent and robust solutions are available to this research community.[26,49] Moreover, there are many limitations and challenges in current application of ML in this fairly young research field, including but not limited to the absence of overarching data resources,[21] limited capability of multimodal data integration[2] and lack of biologically interpretable ML models.[22]

This review comprehensively surveys the state of ML application in spatial proteomics. First, the current spatial proteomic data resources, which lay the foundation for the utilization of ML, are introduced in detail. Second, the role of ML in spatial proteomic data analysis pipeline is systematically described, followed by the concepts of various ML algorithms. Third, examples of successful studies that have harnessed the power of spatial proteomics based on ML methods are presented as well as analytical tools integrating ML methods. Finally, practical problems and limitations in ML-based spatial proteomics research are discussed, and several solutions are proposed. We expect this methodological overview will help promote the development of novel ML methods for spatial proteomics, provide guidance for biologists seeking to apply ML methods and stimulate future research in cell biology.

2. DATA RESOURCES IN SPATIAL PROTEOMICS

2.1. Data Acquisition. Spatial proteomic data acquisition approaches for detecting the localization and abundance of proteins within subcellular compartments can be divided into two categories: quantitative MS-based approaches and fluorescent imaging-based approaches.[21] In general, techniques employing the former type include organelle fractionation and proximity labeling, and the latter type of approaches visualize proteins of interest by using affinity reagents (such as antibodies) or fluorescent protein fusions.[1]

MS-based workflows carry out biochemical fractionation to separate subcellular compartments before MS analysis. The basic strategy to produce discrete organelle fractions is differential centrifugation or density centrifugation according to physical properties of organelles.[27] Single-organelle profiling applies biochemical fractionation methods to produce fractions enriched for a target organelle.[50−52] By contrast, multi-organelle profiling partially separates all subcellular compartments simultaneously and attempts to avoid the error caused by incomplete organelle purification.[18,26] Protein correlation profiling (PCP),[16,29] localization of organelle proteins by isotope tagging (LOPIT)[53,54] and dynamic organellar maps[55,56] are three multi-organelle profiling methods commonly used. Following the fractionation, fractions are processed by quantitative MS to generate protein distribution profile. The principle of these methods is that proteins from the same organelle will share similar distribution profiles across subcellular fractions.[57]

Antibody-mediated affinity purification-mass spectrometry (AP-MS) experiments have been utilized to identify protein−protein interaction (PPI).[58] It is widely recognized that proteins must be adjacent in the same compartment to interact. Proximity labeling techniques are useful for AP-MS experiments.[19,59] There are two main proximity labeling approaches, namely, engineered ascorbate peroxidase and proximity-dependent biotin identification (BioID).[60,61] Following labeling, labeled proteins can be identified by MS analysis to obtain spatial PPI data.[62,63]

Imaging-based spatial proteomics approaches visualize subcellular proteins in situ without the need for cell lysis. Fluorescent labeling of proteins is required prior to microscopic imaging.[39] Immunofluorescence assay based on fluorescently labeled affinity reagents is a commonly used fluorescent imaging approach.[64] Affinity reagents include antibodies,[65] affimers[66] and aptamers.[67] Another fluorescent imaging approach is visualizing proteins by expression of fluorescent protein fusions in living cells, where proteins of interest are genetically modified to express fluorescent tags.[20,68] Gene-editing techniques are employed to insert genes encoding fluorescent proteins at endogenous genomic loci, such as CRISPR-Cas9-based techniques.[69,70] Images of fluorescently labeled proteins are acquired using high-throughput fluorescence microscopy and processed for detecting protein subcellular localization.[71]

2.2. Data Repositories. The above data acquisition techniques give rise to diverse spatial proteomic data. The

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

Table 1. Comprehensive Collection of Spatial Proteomic Data Repositories[a]

Data Repository	Organism	Description	URL
MS-based spatial proteomic data repositories			
SubCellBarCode	*Homo sapiens*	A resource providing the spatial organization of the proteome in five human cancer cell lines, generated by a pipeline consisting of subcellular fractionation and MS-based quantification.[72]	https://www.subcellbarcode.org
PSL-LCCL	*Homo sapiens*	This database quantifies proteins in six cytosolic membrane-bound organelles in the human liver cancer cell line SK_HEP1 using MS.[73]	http://www.igenetics.org.cn/project/PSL-LCCL/
HeLa Spatial Proteome database	*Homo sapiens*	A comprehensive database providing protein subcellular localization and absolute copy number information for over 8700 proteins from HeLa cells, generated by dynamic organellar maps.[55]	http://www.mapofthecell.org
Human Cell Map	*Homo sapiens*	This database presents the subcellular localizations and interactions of 4145 proteins in the human embryonic kidney cell line HEK293 based on BioID and MS analysis.[74]	https://humancellmap.org
BioPlex	*Homo sapiens*	This database visualizes two cell-line-specific PPI networks generated by AP-MS for human HEK293T cells and HCT116 cells and matches proteins with their subcellular fractionation profiles.[75]	https://bioplex.hms.harvard.edu/explorer/
Obesity-induced non-alcoholic fatty liver disease database	*Mus musculus*	Through a MS workflow adapted from PCP, this database determines changes in subcellular distributions of ∼6000 mouse liver proteins during the development of steatosis.[16]	Unavailable
Mouse pluripotent stem cell spatial proteome database	*Mus musculus*	A database that maps more than 5000 proteins to 14 compartments in mouse pluripotent stem cell, generated by the hyperLOPIT approach.[76]	https://lgatto.shinyapps.io/christoforou2015/
Prolocate	*Rattus norvegicus*	This database assigns more than 6000 rat liver proteins to eight subcellular compartments based on subcellular fractionation and quantitative MS and provides quantitative data of each protein.[77]	https://prolocate.cabm.rutgers.edu/
PRoteomics IDEntifications Database (PRIDE)	Multiple organisms	The largest repository for MS data, including raw MS data and processed quantification data of spatial proteomics.[78]	https://www.ebi.ac.uk/pride/
Panorama	Multiple organisms	A knowledge base that stores targeted proteomics data, including raw data and processed data of MS-based spatial proteomics.[79]	https://panoramaweb.org
PeptideAtlas	Multiple organisms	A publicly accessible resource that provides peptides identified in tandem MS proteomics studies such as MS-based spatial proteomics.[80]	http://www.peptideatlas.org
Mass Spectrometry Interactive Virtual Environment (MassIVE)	*Homo sapiens*	A resource that stores human proteomics MS data including raw MS files, identified peak lists and result files from AP-MS experiments.[81]	http://proteomics.ucsd.edu/ProteoSAFe/datasets.jsp
Imaging-based spatial proteomic data repositories			
Yeast GFP Fusion Localization Database (YeastGFP)	*Saccharomyces cerevisiae*	This database classifies 4156 yeast proteins into 22 distinct subcellular localization categories by expressing GFP fusions.[82]	https://yeastgfp.yeastgenome.org
Collection of Yeast Cells Localization Patterns (CYCLoPs)	*Saccharomyces cerevisiae*	This database collects ∼300,000 micrographs, and depicts the localization and abundance dynamics of more than 4000 yeast proteins under different conditions.[46]	http://cyclops.ccbr.utoronto.ca/
Cellbase	*Saccharomyces cerevisiae*	A database describing protein abundance and localization changes in 4085 GFP-tagged yeast strains under six conditions.[68]	http://cellbase.epfl.ch/
Localization and Quantitation Atlas of the yeast proteomE database (LoQAtE)	*Saccharomyces cerevisiae*	This database provides localization and abundance of 5330 GFP-tagged proteins under various growth conditions and genetic backgrounds.[83]	http://www.weizmann.ac.il/molgen/loqate/
YeastRGB	*Saccharomyces cerevisiae*	A database for visualizing and comparing fluorescently tagged protein abundance and localization across yeast cells and strains.[84]	http://www.yeastRGB.org
Database of High Throughput Screening Hits (dHITS)	*Saccharomyces cerevisiae*	A database curating high-throughput screens and displaying the expression and localization of yeast proteins.[85]	https://www.dhitsmayalab.tk/firstPage.php
Yeast Resource Center Public Image Repository (YRC PIR)	*Saccharomyces cerevisiae*	A large database of fluorescence microscopy images depicting the subcellular localization and colocalization of yeast proteins.[86]	https://images.yeastrc.org
Yeast Protein Localization Plus Database (YPL+)	*Saccharomyces cerevisiae*	This database provides 500 sets of image data derived from high-resolution microscopic analyses of GFP-tagged yeast proteins.[87]	https://yplp.yeastgenome.org
Dynamic Proteomics database	*Homo sapiens*	This database represents localization and dynamics following drug addition of 2180 human proteins fused to yellow fluorescent protein in the nonsmall lung cell carcinoma cell line H1299.[88]	Unavailable
Cell Image Library (CIL)	Multiple organisms	A curated repository of cellular images from various organisms.[89]	http://www.cellimagelibrary.org
Image Data Resource (IDR)	Multiple organisms	A repository that collects and integrates bioimage data generated by different imaging approaches, including imaging-based spatial proteomics.[90]	https://idr.openmicroscopy.org
Broad Bioimage Benchmark Collection (BBBC)	Multiple organisms	A publicly available collection of annotated high-throughput microscopy image sets.[91]	https://data.broadinstitute.org/bbbc/

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

Table 1. continued

Data Repository	Organism	Description	URL
Multisource-based spatial proteomic data repositories			
Human Protein Atlas (HPA)	*Homo sapiens*	As part of the HPA, the Cell Atlas maps the localization of 12,003 human proteins to various subcellular structures by using antibody-based immunofluorescence microscopy with validation by MS.[5]	https://www.proteinatlas.org
OpenCell	*Homo sapiens*	A database that maps the subcellular localization and interactions of 1310 fluorescently tagged proteins in human HEK293T cells by performing immunopurification-MS and live-cell imaging.[92]	https://opencell.czbiohub.org
Universal Protein Resource (UniProt)	Multiple organisms	A comprehensive database that provides complete protein information, including the subcellular localization extracted from literatures and computational analyses.[93]	https://www.uniprot.org
Gene Ontology (GO)	Multiple organisms	This resource provides knowledge regarding functions of genes and gene products, and the cellular compartment namespace in GO is essential for the annotation of protein localization.[94]	http://geneontology.org
eukaryotic Subcellular Localization DataBase (eSLDB)	Multiple organisms	A database that collects experimentally determined and predicted subcellular localization of proteins from five eukaryotic organisms.[95]	http://gpcr2.biocomp.unibo.it/esldb
COMPARTMENTS	Multiple organisms	This database integrates experimental data and predictions to hold subcellular localization information for 22,705 human proteins, 6696 yeast proteins, as well as proteins of other eukaryotes.[96]	http://compartments.jensenlab.org
LocDB	*Homo sapiens* and *Arabidopsis thaliana*	A manually curated database with experimental annotations and predictions for the subcellular localizations of 13,342 human proteins and 6262 proteins of *Arabidopsis thaliana*.[97]	https://www.rostlab.org/services/locDB/
ComPPI	Multiple organisms	A manually curated database that provides subcellular compartment-specific PPI and protein subcellular localization in four species.[98]	http://comppi.linkgroup.hu/
CellWhere	Multiple organisms	An integrated resource that provides the local interaction network in subcellular compartments.[99]	https://www.sys-myo.com/cellwhere/
SUBcellular location database for *Arabidopsis* proteins (SUBA4)	*Arabidopsis thaliana*	A comprehensive collection of manually curated data of MS-based subcellular proteomics, imaging-based subcellular proteomics, PPI and predicted protein localization.[100]	https://suba.live/
Translocatome	*Homo sapiens*	A manually curated database that collects 213 human translocating proteins with extensive information (experimental validation, translocation mechanism, local PPI, etc.).[101]	http://translocatome.linkgroup.hu/
PeroxisomeDB	Multiple organisms	A curated database organizing information on 2819 peroxisomal proteins from 38 organisms and their interactions.[102]	http://www.peroxisomedb.org
MitoCarta	*Homo sapiens* and *Mus musculus*	This database provides human and mouse mitochondrial proteins along with annotations of their submitochondrial localization based on literature review and prediction.[103]	http://www.broadinstitute.org/mitocarta/
MitoMiner	Multiple organisms	This database incorporates mitochondrial localization data from a variety of resources, including MS-based and imaging-based experimental data, as well as computational prediction.[104]	Unavailable
RareLSD	*Homo sapiens*	A manually curated database that provides complete information on 63 human lysosomal enzymes associated with rare diseases.[105]	https://webs.iiitd.edu.in/raghava/rarelsd/

[a]The URLs of currently available data repositories are provided. MS, mass spectrometry; BioID, proximity-dependent biotin identification; AP-MS, affinity purification-mass spectrometry; PCP, protein correlation profiling; PPI, protein−protein interaction; GFP, green fluorescent protein.

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

expected data output from these techniques consists of three main forms: raw data, processed quantitative data and qualitative data for protein localization. In MS-based techniques, the raw output is tandem mass spectral data and needs to be analyzed by quantification measurement tools. The quantification result is a list of proteins alongside their quantitation in multiple fractions, which can be further analyzed to generate the subcellular localizations of proteins. Imaging-based spatial proteomic techniques produce raw images of compartment localization of proteins. These images can be directly analyzed or used to extract quantitative features (such as shape and fluorescence intensity) to obtain protein localization information.[21] Here, we provide a comprehensive overview of spatial proteomic data repositories (as shown in Table 1).

MS-based spatial proteomic approaches have been applied to various cell lines from different organisms, yielding large-scale high-resolution organellar maps of these cells. A robust MS-based analysis pipeline has been developed to generate subcellular maps of 12,418 individual proteins across five human cell lines, and spatial organization of the proteome in these cells was used to construct the SubCellBarCode database.[72] Some databases contain both the subcellular localization and abundance information on proteins. Among them, the PSL-LCCL database quantified proteins in six membrane-bound organelles in the human liver cancer cell line SK_HEP1.[73] The absolute protein quantification (also known as protein copy numbers) can be captured by dynamic organellar maps. Based on dynamic organellar maps, the HeLa Spatial Proteome database provides localization and absolute copy number information for over 8700 proteins from HeLa cells.[55] There are also several databases incorporating intracellular localization and cell-line-specific PPI data obtained by MS-based approaches, such as the Human Cell Map[74] and BioPlex.[75] In addition to human, these advanced MS-based approaches have been applied to investigate the spatial proteome in other organisms.[18,106] For example, researchers who constructed the obesity-induced non-alcoholic fatty liver disease (NAFLD) database have used the PCP technique to monitor levels and subcellular distributions of approximately 6000 mouse liver proteins during development of NAFLD.[16] The hyperLOPIT,[107] an extension of LOPIT, has been applied to mouse pluripotent stem cell, resulting in the mapping of more than 5000 proteins to 14 compartments.[76] The Prolocate database assigns more than 6000 rat liver proteins to eight subcellular compartments and provides quantitative data of each protein.[77] In some cases, the spatial proteomic data obtained are only provided in the form of data set without being submitted to a specific database, and sharing data through supplementary data files has been common. For example, 11 subcellular proteomic data sets were processed and the identified proteins along with normalized MS signal values were provided in the supporting information of a tool development study.[108] Besides, some MS-based spatial proteomic data sets have been packaged and used as benchmarks for related studies. These data sets include the assignment of 527 proteins to different compartments in *Arabidopsis thaliana*[31] and 329 proteins in *Drosophila melanogaster* based on LOPIT technique.[109] Both data sets are integrated in pRolocdata package, which contains a total of 16 spatial proteomic data sets from LOPIT and PCP.[110]

High-thought imaging-based spatial proteomic approaches make large-scale cellular image data sets available for research community. The LiveCellNet data set, integrated in DeepCell data set, provides a large number of cellular images obtained by fluorescence microscopy.[111] As a model organism, *Saccharomyces cerevisiae* has been widely used in the study of eukaryotic biology, and numerous image data sets of yeast strains have been created in the past decade. For instance, the localizations for over 200 of green fluorescent protein (GFP) tagged proteins of *Saccharomyces cerevisiae* were determined by colocalizing them with seven known markers of endomembrane compartments.[112] Several databases have been constructed to provide the global organellar mapping of proteins in *Saccharomyces cerevisiae*, such as the Yeast GFP Fusion Localization Database (YeastGFP) classifying 4156 yeast proteins into 22 distinct subcellular localizations.[82] Imaging-based approaches were used to depict the localizations and abundance dynamics of proteins under different conditions or genetic mutations, as these approaches have the advantage of detecting protein localization in living cells. Owing to these systematic imaging studies using fluorescent protein fusions, some databases describing protein quantitation and localization variations under conditions or yeast strains are now available, including Collection of Yeast Cells Localization Patterns (CYCLoPs),[46] Cellbase,[68] Localization and Quantitation Atlas of the yeast proteomE database (LoQAtE)[83] and YeastRGB.[84] Furthermore, a variety of databases have collected microscopic imaging data of protein localization patterns in the yeast *Saccharomyces cerevisiae* from separate experiments, such as the Database of High Throughput Screening Hits (dHITS),[85] Yeast Resource Center Public Image Repository (YRC PIR)[86] and Yeast Protein Localization Plus Database (YPL+).[87] The fluorescent protein fusion approaches have also been carried out to study the spatial organization of human proteins. For example, the localization and dynamics following drug addition of 2180 human proteins fused to yellow fluorescent protein were obtained using time-lapse fluorescence microscopy and displayed in the Dynamic Proteomics database.[88] To compare the reliability of immunofluorescence and fluorescent protein fusion, a comparative study was conducted to systematically analyze the localizations of more than 500 human proteins in Vero or HeLa cells, which were obtained by these two imaging-based approaches, respectively.[65] The results proved both approaches to share high reliability, and the integration was an effective strategy. Recently, projects aiming to comprehensively map the human body at single-cell resolution have been sponsored, such as the Human Biomolecular Atlas Program (HuBMAP), which supports technology development and data acquisition.[113] A CO-Detection by indEXing (CODEX)[114] spatial proteomic image data set from HuBMAP has been used to test the usefulness of an analysis tool.[115]

MS-based and imaging-based approaches give complementary insights into spatial proteome, and the combination of both approaches has demonstrated strong synergy.[116] As an integrated part of the Human Protein Atlas (HPA), the Cell Atlas maps the localization of 12,003 human proteins from a variety of cell lines into 30 subcellular structures. The Cell Atlas data were generated by antibody-based immunofluorescence microscopy and validated by the MS-based hyperLOPIT approach.[5] Combining these two approaches enables a more complete coverage of the organelle proteome, such as the mitochondrial proteome.[60,117] This combination strategy has also been utilized to characterize the physical interactions of proteins in different species, such as *Chlamydomonas*

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

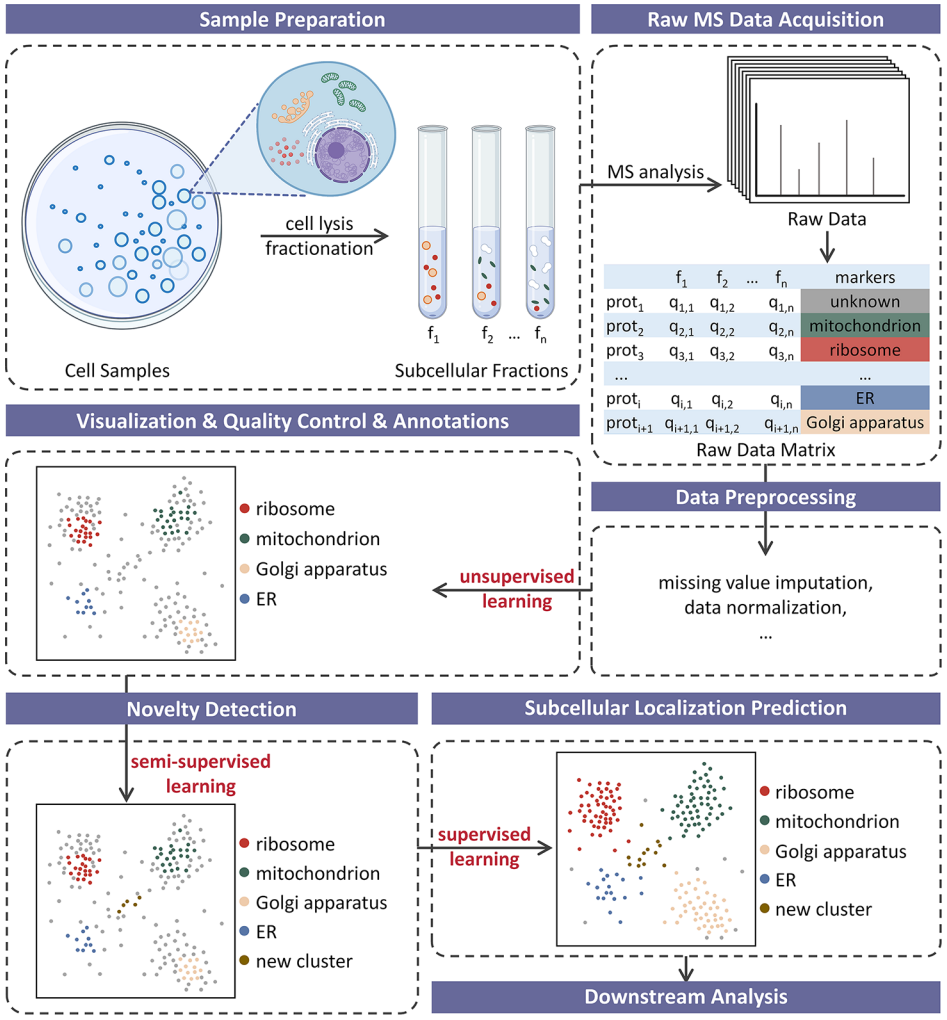


Figure 1. Roles of different ML methods in MS-based spatial proteomic data analysis pipeline. ER, endoplasmic reticulum.

reinhardtii.[118] To enable the comprehensive mapping of protein localization and interactions in human embryonic kidney (HEK) 293T cell, a library of 1310 fluorescently tagged HEK293T cell lines was constructed, followed by immunopurification-MS and imaging. The generated data were shared at OpenCell.[92] Furthermore, the most significant advantage of the combination strategy is the ability to provide spatially and temporally resolved proteomic maps simultaneously. By integrating MS-based proteomics and living cell microscopy, spatial and temporal alterations in human primary fibroblasts proteome during human cytomegalovirus (HCMV) infection were determined.[119] These important data obtained by such multidisciplinary analysis allow researchers to understand organelle regulation during various cellular processes.

Raw data from spatial proteomic studies are also valuable data resources. Various large public repositories have been created to allow raw data from different experiments to be shared and reanalyzed. In general, submitting raw data to suitable data repositories is a prerequisite for publication of spatial proteomic research. In MS-based research, raw mass spectra files and associated peak lists can be submitted to some public repositories,[120,121] such as the PRoteomics IDEntifications Database (PRIDE),[78] Panorama,[79] PeptideAtlas[80] and Mass Spectrometry Interactive Virtual Environment (MassIVE).[81] For imaging-based spatial proteomics, the raw microscopic image data can be deposited in public repositories such as the Cell Image Library (CIL),[89] Image Data Resource (IDR)[90] and Broad Bioimage Benchmark Collection (BBBC).[91] Before these public repositories became accessible, researchers shared the raw data through their institutional servers, making it difficult and impractical for subsequent researchers to obtain comprehensive data. The establishment of these large data repositories ensures data accessibility and reusability for ML-based spatial proteomic studies.

2.3. Data Resources for Organelle Markers. An organelle marker is a known protein that resides in a specific subcellular compartment in cells and conditions of interest.[26] As the initial step of the spatial proteomic data analysis workflow, the vigorous selection of markers is essential to ensure the high accuracy of eventually obtained localizations of the proteins. In MS-based spatial proteomics, the abundance distribution of organelle markers can be used to train ML models to determine the localizations of unknown proteins.[55] Non-marker proteins are assigned to organelles according to the similarity of their gradient distributions to those of organelle markers. The SubCellBarCode database and the HeLa Spatial Proteome database were constructed based on the spatial profiles of various marker proteins in different cell lines.[55,72] In the analysis pipeline of imaging-based approaches, the selection of markers is also an important step. Cells or subcellular structures are required to be segmented by measuring the signal intensities for structure-specific markers,

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

such as cell boundary markers, cytosolic markers and nucleolar markers.[122−124]

There are many protein mapping data resources that can provide information on marker proteins for different species, cell lines or organelles. The Universal Protein Resource (UniProt)[93] and the cellular compartment namespace in Gene Ontology (GO)[94] are essential resources for protein annotation and marker definition. Both resources are commonly used to determine reference markers of many organisms in spatial proteomic studies. There are also some databases collecting experimentally determined and predicted subcellular localizations of proteins. Most of these databases provide the spatial information on proteome in multiple organisms and help in marker selection for respective species, including the eukaryotic Subcellular Localization DataBase (eSLDB),[95] COMPARTMENTS[96] and LocDB.[97] Specifically, the COMPARTMENTS database integrates manually curated annotations, experimental data and predictions of protein localization from biomedical literatures. It maps subcellular localizations for 22,705 human and 6696 yeast proteins and also covers other eukaryotes such as mouse, fruit fly and *Caenorhabditis elegans*.[96] In addition, some databases contain both protein localization and PPI information manually curated from numerous experimental literatures, such as the ComPPI,[98] CellWhere[99] and SUBcellular location database for *Arabidopsis* proteins (SUBA4).[100] The ComPPI and CellWhere provide the information on subcellular compartment-specific PPI and protein subcellular localization for many organisms.

It is worth noting that since proteins can reside in multiple subcellular structures simultaneously and their localizations are dynamic, the validation of organelle markers under different conditions is crucial. The Translocatome database is a dedicated collection of 213 human translocating proteins with extensive information (experimental validation, translocation mechanism, local PPI, etc.), which provides a reference for the selection of marker under different conditions.[101] Besides, some manually curated databases collate proteins localized to specific organelles from a variety of resources, giving new insights into the definition and extension of organelle markers. For example, the PeroxisomeDB (peroxisomal proteome),[102] MitoCarta (human and mouse mitochondrial proteins),[103] MitoMiner (mitochondrial proteins in mammals, zebrafish and yeasts)[104] and RareLSD (human lysosomal enzymes)[105] are customized for different organelles in various organisms. In other studies, the marker sets were included in supporting information to ensure data reproducibility and provide marker information for related studies.[26,108,125] These data resources serve as references for the curation of reliable sets of organelle markers.

3. HOW ML IS INTEGRATED INTO SPATIAL PROTEOMICS

So far, various ML methods have been integrated into the data analysis pipeline of spatial proteomics. The appropriate selection and execution of ML methods can ensure precise inference of proteome-wide localization from the data. Due to the differences in data types, the data analysis pipelines of MS-based and imaging-based spatial proteomics are quite different, hence the varied applications of ML methods in these pipelines. Here, we will elaborate these two analysis pipelines and explain the crucial roles of various ML algorithms in the process, followed by their concepts.

3.1. MS-Based Spatial Proteomic Data Analysis Pipeline.

The data analysis pipeline of MS-based spatial proteomics is summarized in Figure 1. After sample preparation and raw mass spectral data acquisition, some quantification tools (such as the MaxQuant)[126] are required for statistically robust protein/peptide identification and quantification. Quantitative measurement generates a corresponding quantitative data matrix for each sample, where each row represents a protein and each column represents a gradient fraction. As prior knowledge, some proteins are organelle markers with known subcellular localization, and the localization annotation of these markers in the data matrix is vital for subsequent analysis. Data preprocessing is needed for the obtained data matrix, mainly including missing value imputation and data normalization. The significant impact of missing value imputation on the downstream analysis of spatial proteomic data has been investigated.[26] Several missing value imputation methods are based on ML algorithm, such as the k-nearest neighbors (kNN) imputation.[127,128] In addition, many data normalization approaches can be used for adjusting the distribution of spatial proteomic data to remove unwanted variations,[129] such as the variance stabilization normalization[130,131] and EigenMS.[132]

Following data preprocessing, visualization of the full preprocessed data set in one figure is a crucial step in the analysis pipeline.[23] Unsupervised dimensionality reduction and clustering are two commonly used visualization methods. Based on the data visualization results, the quality of the data can be further assessed and controlled.[26] The use of dimensionality reduction and clustering methods, such as principal component analysis (PCA),[133] t-distributed stochastic neighbor embedding (t-SNE)[73] and hierarchical clustering,[134] prove to be an efficient approach for MS-based spatial proteomic data quality control and checking organelle separation. Based on the assumption that proteins sharing the same subcellular localization should be similarly distributed across gradient fractions, the data of these colocalized proteins should form well-defined structures in the visualization result.[135] Therefore, the data quality can be inspected by observing the data structures. Intuitively, the overall structure of data can reflect the data quality, that is, the existence of clusters in the data is the prerequisite for determining the classification boundaries separating the subcellular compartments. Another assessment for quality control is to overlay organelle markers in the visualization result (such as the PCA plot) and observe whether an expected organelle-related cluster is present and concentrated rather than widespread. Similarly, the quality of clusters can also be evaluated by checking whether known organelle markers appear in the corresponding clusters.[26]

After data visualization and clustering, the clusters are annotated with specific subcellular compartments based on these marker proteins. Enrichment analysis using data resources such as UniProt and GO provides the identity of each cluster.[72] However, owing to the limited number of organelle markers provided in these relevant databases, it is difficult to comprehensively collect marker proteins covering all the subcellular compartments, making it tough to find all the organelle-related clusters.[125] Thus, for those proteins localized in organelles without any suitable markers, the subsequent localization prediction will be completely mistaken because they will be forcibly assigned to other organelles. To address this issue, some novelty detection algorithms have

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

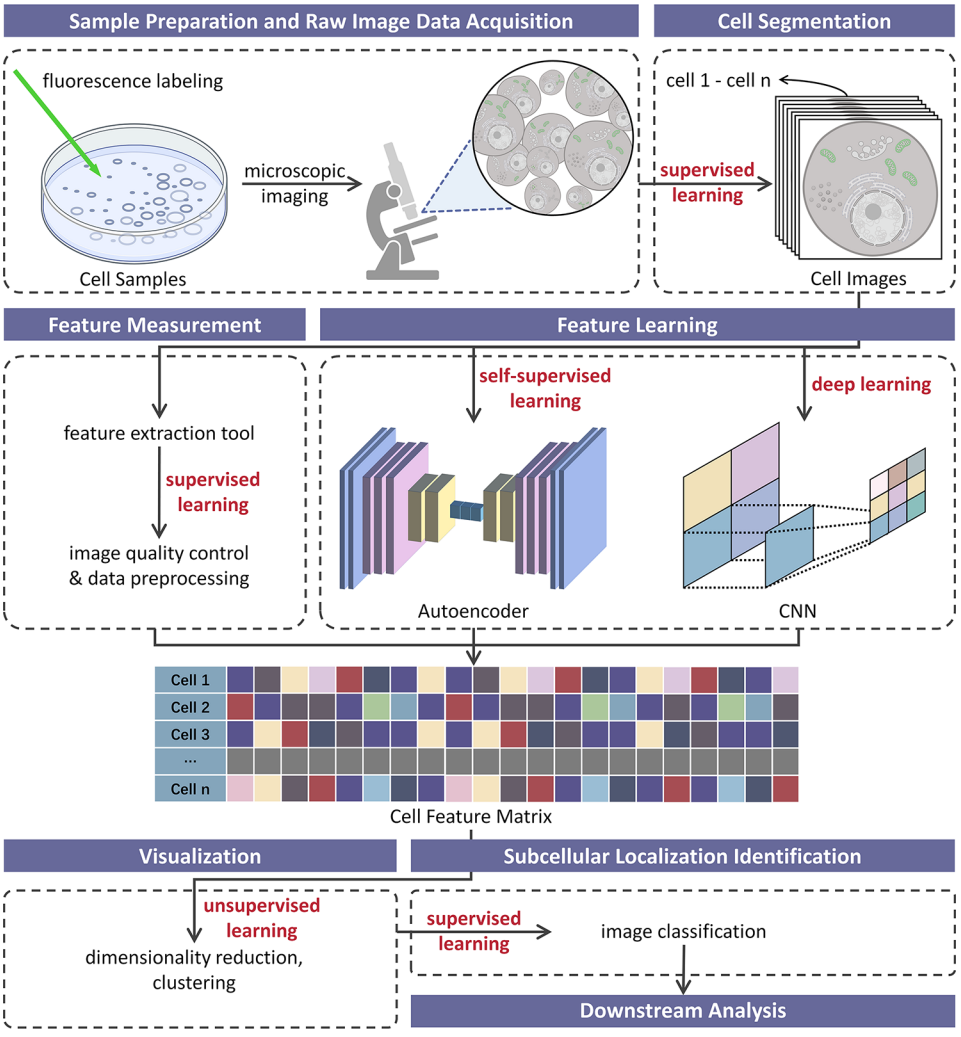


Figure 2. Roles of different ML methods in imaging-based spatial proteomic data analysis pipeline. CNN, convolutional neural network.

been developed to identify new subcellular clusters (including organelles and protein complexes) in addition to clusters annotated by known markers. The aim of novelty detection is to find new clusters from unlabeled data according to labeled markers, which matches the concept and aim of semi-supervised ML algorithms. Therefore, the semi-supervised ML algorithms are useful for novelty detection to effectively detect putative protein groupings in MS-based spatial proteomics data.[136]

According to the distribution profiles of marker proteins in gradient fractions, the subcellular localization prediction of non-marker proteins can be performed by constructing ML models based on supervised learning algorithms. The gradient profiles of markers are used as training and test data to train a ML model for the prediction of subcellular localization. The trained model is then applied to assign non-marker proteins to subcellular localizations by mapping their gradient profiles to organelles. Various supervised ML algorithms have been used for this classification task, such as the support vector machine (SVM),[137] random forest (RF),[37] naive Bayesian (NB) classifier,[138] kNN[43] and neural networks.[119] While the reliability of prediction results largely depends on the quality of spatial proteomic data, it is advisable to compare different ML classification algorithms and guarantee optimal application of the algorithm, since the diverse principles of these supervised ML algorithms may lead to inconsistent results.[38,110] When the number of subcellular compartments is more than two, the spatial assignment of proteins is essentially a multi-class classification task. It should be noted that some proteins are not localized in a single compartment, instead, around 50% of proteins are found to reside in multiple subcellular compartments.[5] In this case, it is necessary to construct a multi-label classifier to achieve multi-localization prediction with classification probability scores. The quality of predictions can be further assessed by comparison with previously published data sets of protein localization. Downstream analysis of the predicted qualitative protein localization data can yield biological or medical knowledge of interest.

3.2. Imaging-Based Spatial Proteomic Data Analysis Pipeline. The general workflow of imaging-based spatial proteomic data analysis is illustrated in Figure 2. The raw data obtained by imaging-based spatial proteomic approaches are microscopic images with a wide field of view, and each image contains hundreds of individual cells. Typically, each cell in these images needs to be identified and analyzed individually, and this process is known as cell segmentation.[139] Cell segmentation is a crucial step and aims to distinguish each single cell from other cells and the background. It can be realized by building a dedicated ML model. Specifically, all pixels in an image can be grouped into three classes: cell

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

boundary, cell interior and background pixel.[49] In this way, cell segmentation is effectively converted into an image classification task. Once the manually annotated images are collected, a pixel classifier can be trained to find optimal segmentation boundaries and ultimately achieve cell segmentation by distinguishing between three classes of pixels.[140] Currently, various methods have been developed for cell segmentation, and they generally fall into two categories: nuclear segmentation and whole-cell segmentation. Supervised ML algorithms are often used to construct such pixel classifiers, for instance, the Ilastik tool applied a RF classifier to identify boundaries between neighboring cells.[141]

Like MS-based spatial proteomic data, image data from imaging-based approaches can also be transformed to a data matrix with quantitative values. This step conducts feature measurements for each segmented cell image. As a result, each cell is represented as a vector. These features capture the protein distribution patterns related to cellular morphology and describe the quantitative information on each cell, including shape, intensity, texture and context of cell images.[49] Several image processing tools are capable of extracting features for cell images, such as the ImageJ software.[142] These image features also need to be preprocessed. The preprocessing of extracted cell features is a necessary step to enhance the recognition of underlying localization patterns. Missing value imputation and data normalization are two main steps in data preprocessing, and the kNN imputation has also been used to reconstruct missing values in image features.[30]

Due to possible errors in sample preparation and image processing, some cell images may be of poor quality. These outliers include out-of-focus cells, partly visible cells and inappropriately segmented cells, and image quality control is an important step to detect and remove these outlier cells.[143] The application of supervised ML algorithms is an effective way to identify problematic images. This strategy aims to build a ML model which can discriminate between normal and outlier cells.[144] By providing examples of outliers and their extracted features, a supervised ML classifier can be trained to accurately detect outliers.[49] Various supervised ML algorithms have been used for cell image quality control, for example, the CellClassifier tool can detect outlier cells based on the SVM model,[145] and a RF model has been trained to remove non-cell images.[146]

Some of the measured features used for analysis are less informative or redundant and should be filtered or merged. Thus, the original feature set can be optimized by dimensionality reduction methods. Feature selection is a method to reduce the feature dimensionality by selecting the best feature subset, and there are many kinds of feature selection algorithms.[147−149] Among them, the support-vector-machine-based recursive-feature elimination (SVM-RFE) method has been applied in imaging-based spatial proteomic data analysis.[150] The SVM-RFE algorithm iteratively removes features according to the feature weights from a SVM classifier until the model achieves the highest performance. In addition, unsupervised dimensionality reduction methods used in MS-based spatial proteomics are also useful to reduce the dimensionality of image features, such as PCA. PCA can be used to visualize image features by mapping them into two or three dimensions.[30] Similarly, unsupervised clustering methods (such as the K-means clustering[151] and hierarchical clustering[152]) can also be used for identification of cell subpopulation and pattern recognition of protein localization based on image features. These informative features are further used as inputs to train supervised ML models (such as SVM[153] and RF[146]) for the prediction of proteins subcellular localizations in cell images or identification of cell phenotypes.

With the recent dominance of DL in computer vision tasks, a revolution of imaging-based spatial proteomic data analysis is underway.[154] The main advantage of DL methods in analyzing cell images is that they can directly learn image patterns without any feature engineering. At present, the most commonly used DL model for image analysis is the convolutional neural network (CNN), which has proved to be powerful in biomedical image segmentation and classification.[155] The CNN is suitable for cellular morphological profiling and can be applied for cell segmentation in spatial proteomic images.[156] For example, the CellProfiler tool can be configured to make use of CNN to segment cells.[157] Moreover, these DL models can learn and extract useful features automatically from microscopic images using nonlinear transformations, hence their superior performance on protein localization prediction compared with classical feature measurements.[155]

In addition, DL-based self-supervised learning models can automatically learn image features without any image labels and have been used in cell segmentation and automatic representation of cell images. Deep autoencoder (AE) is a typical self-supervised technique which learn the latent image features based on the image itself. As a typical AE, the U-Net performs single-cell segmentation end to end and functions as a plugin in the ImageJ software.[158] In order to analyze the image features obtained by DL models or self-supervised learning models, dimensionality reduction methods such as t-SNE and uniform manifold approximation and projection (UMAP) have been used for feature visualization.[155,159] It has been reported that features generated by AE may have better performance than classical features in downstream analysis.[146] Self-supervised learning has become an effective and popular strategy for spatial proteomic image analysis.

3.3. ML Algorithms Adopted in Spatial Proteomics.

As mentioned above, different types of ML algorithms play different roles in spatial proteomic analysis pipeline. Understanding the concepts of various types of ML strategies is the key to the success application of ML in spatial proteomics.

Supervised learning is a mode of ML that projects input features to annotated data labels. It has been used in several early studies to predict localization based on protein sequence.[24] In MS-based and imaging-based spatial proteomic data preprocessing, the kNN is often used to impute missing values.[23] In MS-based spatial proteomics, supervised ML algorithms are commonly employed to build classifiers that associate unannotated proteins to specific subcellular compartments, such as the SVM,[160] kNN,[43] RF[37] and NB.[138] In imaging-based spatial proteomics, these supervised ML algorithms are used in multiple data analysis steps, including cell segmentation, image quality control and identification of protein localization based on image features.[141,145,151,153] The principles of these popular algorithms are quite different and have been introduced elsewhere.[161−164]

DL methods use interconnected layers of nonlinear transformation units to learn from data without labor-intensive feature engineering.[154,165] DL models such as deep neural network (DNN) and CNN have been employed to perform cell segmentation, image representation and protein localization prediction, and their principles were described

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

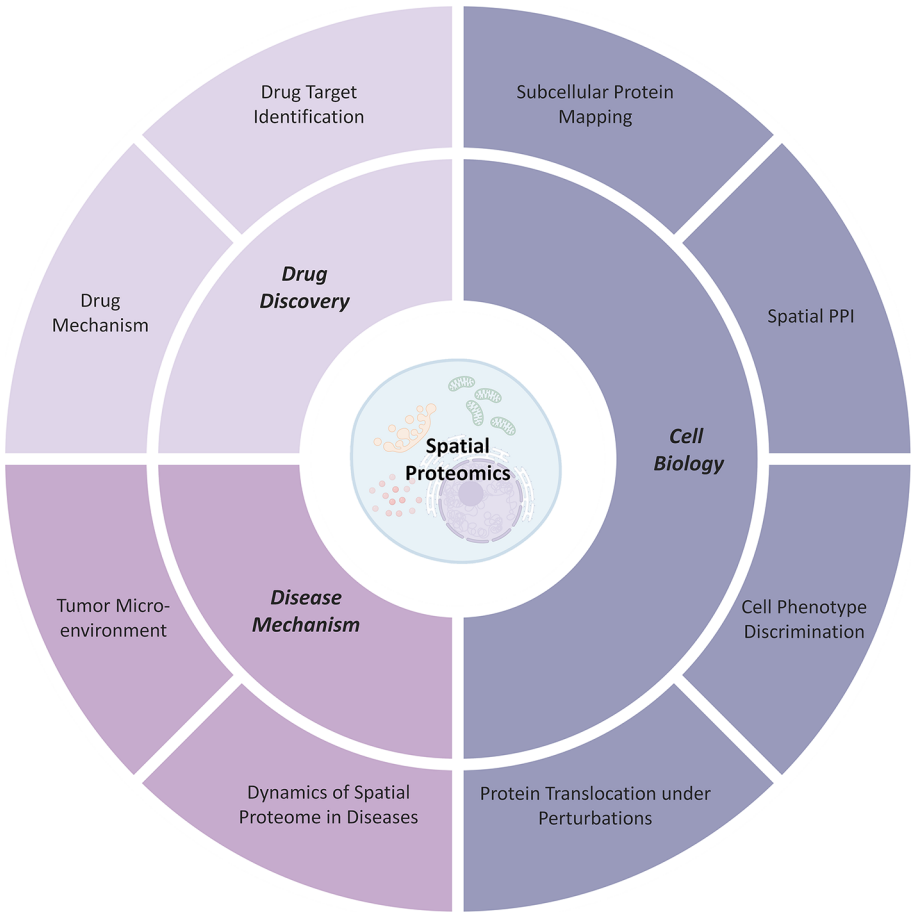


Figure 3. Summary of ML-aided spatial proteomics applications.

elsewhere.[140,154,155,166−169] Different from supervised learning, self-supervised learning is a mode that generates labels from data itself for training. AE is a type of semi-supervised learning method which embeds high-dimensional data into a low-dimensional latent space while preserving the original information on inputs.[170] An AE is composed of two modules: an encoder and a decoder. The encoder maps the input to a hidden representation, and the decoder then decodes the hidden representation to reconstruct the input. It can reduce cell images to feature vectors in a mirror-symmetric paradigm.[171] CNN-based AEs have been widely applied in cell segmentation and protein localization profiling.[171,172] The U-Net is a typical convolutional AE originally developed for cell detection and shape measurement.[158]

Unsupervised learning does not require additional label annotation.[33,173] Unsupervised ML methods play significant roles in data visualization, quality control, feature dimensionality reduction and subpopulation identification.[30,138,155,174] Dimensionality reduction methods enable compression of high-dimensional data into a set of two or three dimensions with maximum retention of initial information.[175] Popular dimensionality reduction methods are implemented by linear or nonlinear transformation, the former including PCA[133] and non-negative matrix factorization (NMF),[176] the latter including t-SNE[177] and UMAP.[178] Clustering is also an unsupervised ML technique which can group data according to the similarity of features.[179] Particularly, subcellular resolution can be assessed by clustering algorithms since proteins with the same subcellular localizations should cluster together. K-means clustering and hierarchical clustering are two commonly employed algorithms and have been reviewed elsewhere.[134,180,181] The most notable characteristic of semi-supervised learning is its ability to utilize both labeled and unlabeled data.[44] Semi-supervised ML methods can be applied for novelty detection in MS-based spatial proteomics. The *phenoDisco* algorithm is semi-supervised and performs iterative cluster merging based on the Gaussian mixture model (GMM) and outlier detection.[110,125] Other semi-supervised ML methods are also available for novelty detection, such as the semi-supervised Bayesian algorithm.[136]

Transfer learning is a paradigm that extracts complementary information from auxiliary data to help solve the primary task.[160] In MS-based spatial proteomics, the transfer learning strategy has been used to integrate heterogeneous data sources to help with the assignment of proteins to subcellular compartments.[43] In imaging-based spatial proteomic studies, transfer learning can be performed by pretraining a DL model and transferring it to the primary task.[40,146] The combination of multiple classifiers can significantly improve the accuracy of cell image classification, which demonstrates the effectiveness of ensemble learning in spatial proteomics.[46,150,182] In addition, multiple instance learning is also a helpful strategy to boost the classification accuracy of microscopic images with whole image level labels.[183,184] These state-of-the-art ML strategies promote the in-depth analysis of spatial proteomic data.

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

Table 2. Recommended Tools for Spatial Proteomic Data Analysis That Integrate ML Methods[a]

Analytical Tool	Description	URL
Analytical tools for MS-based spatial proteomic data		
MSnbase	MSnbase is an R package used for MS data visualization that can conduct raw data import, data preprocessing (such as kNN missing data imputation) and quantitation.[188]	http://www.bioconductor.org/packages/release/bioc/html/MSnbase.html/
pRoloc	pRoloc is an integral analytical tool that provides various functions for spatial proteomic data exploration, including data visualization, quality control, novelty detection and classification, and it provides multiple supervised and unsupervised ML methods.[110]	http://www.bioconductor.org/packages/release/bioc/html/pRoloc.html
pRolocGUI	pRolocGUI is an R package used for visualizing protein subcellular classification by conducting PCA and supervised ML algorithms.[107]	http://www.bioconductor.org/packages/release/bioc/html/pRolocGUI.html
MetaMass	MetaMass utilizes K-means clustering to identify protein groups and maps the subcellular localizations of proteins.[108]	https://github.com/stuchly/MetaMass
TRANSPIRE	TRANSPIRE uses organelle markers to produce synthetic translocation profiles and trains a stochastic variational Gaussian process classifier for predicting protein translocation.[189]	https://github.com/cristealab/TRANSPIRE_JASMS2020
Analytical tools for imaging-based spatial proteomic data		
Ilastik	Ilastik can conduct cell segmentation based on a RF classifier and can also be used for image classification.[141]	https://www.ilastik.org/
ImageJ	ImageJ software can extract image features and provide ML-based plugins to segment cells, such as U-Net.[142]	https://imagej.net/develop/
Cellpose	Cellpose performs precise whole-cell segmentation within broad range of cell images using an optimized U-Net model, without requirement of model retraining.[172]	http://www.cellpose.org/
Squidpy	Squidpy implements numerous DL-based methods for cell segmentation and features extraction.[190]	https://github.com/scverse/squidpy
Cytokit	Cytokit is a comprehensive toolkit for image data analysis which conducts image feature extraction and nuclear segmentation based on DL models.[191]	https://github.com/hammerlab/cytokit
CellProfiler	CellProfiler is a powerful tool for cell image analysis that operates CNN-based cell segmentation, feature extraction and image classification.[157]	https://github.com/CellProfiler
DeepCell Kiosk	DeepCell Kiosk enables the processing of large-scale image data sets by utilizing pretrained DL methods.[111]	https://deepcell.org/
Giotto	Giotto is a toolbox for spatial proteomic image analysis and visualization, and it provides unsupervised dimensionality reduction and clustering algorithms to identify protein localization patterns.[192]	https://rubd.github.io/Giotto_site/
Cellar	Cellar provides numerous unsupervised and semi-supervised ML methods to conduct visualization and comparison of spatial proteomic image data.[115]	https://cellar.cmu.hubmapconsortium.org/app/cellar

[a]All the analytical tools presented are currently available. MS, mass spectrometry; kNN, k-nearest neighbors; PCA, principal component analysis; RF, random forest; DL, deep learning; CNN, convolutional neural network.

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

4. RECENT STATE-OF-THE-ART APPLICATIONS OF ML TO SPATIAL PROTEOMICS

Spatial proteomics have harnessed the power of ML to broaden the range of its applications. As shown in Figure 3, the recent state-of-the-art applications of spatial proteomics fall into the following three categories: (1) cell biology:[122,185] investigation of subcellular protein mapping and spatial PPI, identification of protein translocation under environmental or genetic perturbations, and discrimination of cell phenotypes such as different cell-cycle states; (2) disease mechanism:[119,186] detection of dynamics of spatial proteome in various diseases and analysis of tumor microenvironment based on cell interactions; and (3) drug discovery:[56,187] identification of putative drug target and investigation of drug mechanism according to drug response. In this section, we will present some successful applications of spatial proteomics based on ML methods and introduce several popular analytical tools for spatial proteomic data that integrate ML methods. The descriptions of some recommended analytical tools integrating ML methods are summarized in Table 2.

4.1. MS-Based Spatial Proteomic Applications. In order to comprehensively capture the cell biological characteristics in different species or cell types, various ML methods have been used to predict protein localization from MS-based spatial proteomic data. For example, RF has been applied to comprehensively identify the proteins localized in mitotic chromosomes of chicken DT40 cells,[37] and SVM has been used to assign 8000 proteins to subcellular localizations in mouse primary neurons and human liver cancer cell line SK_HEP1.[18,73] ML methods are also useful to analyze spatial PPI data. NMF has been employed to analyze the BioID data and define the subcellular localizations of 4145 proteins and their interactions in HEK293 cells.[74] Some comparative studies have investigated the drug response based on changes in proteomic subcellular profiles of control and drug-treated cells. For instance, dynamic changes in organelle profiles between control and drug-treated mouse macrophage cells have been detected by comparison PCA.[193] Another study used SVM to predict protein-organelle associations in dendritic cells and identified protein translocations after treatment with approved drugs such as tamoxifen and prazosin.[56] These comparative studies help reveal drug mechanisms. MS-based spatial proteomics can also reveal disease mechanisms and facilitate drug discovery. To determine spatial and temporal changes in subcellular proteome in human primary fibroblasts during HCMV infection, various supervised ML algorithms (including SVM, RF and neural networks) have been used to predict protein localization.[119] The results demonstrated that the translocation of unconventional myosin MYO18A was necessary in HCMV replication, and it was identified as a potential anti-HCMV target.

To obtain robust predictions of protein localizations, some efforts have been made to develop specialized MS-based spatial proteomic analysis strategies. A robust MS-based analysis pipeline has been developed to generate the delineation of subcellular localizations of proteins in five cell lines.[72] Specifically, this pipeline performed quality control by PCA to examine the resolution of subcellular compartments based on fractionation profiles. Then, the t-SNE method was used to visualize and identify marker proteins clusters, and these clusters were annotated with specific subcellular compartments using GO and UniProt annotations. Finally, marker proteins were divided into a training dataset and a test dataset to construct a SVM classifier. Based on this pipeline, multiple condition-dependent protein localization analyses were conducted, including the cell-type-specific protein localization, effects of splicing and protein domains on localization and protein relocalization after EGFR inhibition.[72] Novelty detection is an effective step to improve the accuracy of protein localization prediction. The ability of the *phenoDisco* algorithm and the semi-supervised Bayesian approach in discovering potential protein groupings have been demonstrated across multiple types of MS-based spatial proteomic data sets.[125,136]

Although a variety of supervised ML algorithms have been successfully employed to build classifiers that associate proteins to specific subcellular compartments, these methods cannot quantify the uncertainty in the assignment of proteins to multiple subcellular localizations. To quantify this uncertainty, a Bayesian generative classifier based on GMM has been created to generate probability distributions for the subcellular localizations of proteins.[194] Compared with various ML methods, this uncertainty quantification approach performs competitively on 19 MS-based spatial proteomic data sets and draws meaningful biological results in several cases. Moreover, in order to improve the quality and quantity of protein assignments, some studies combine multiple data sets from experiments resolving different subcellular compartments. This data combination strategy has been applied in a spatial proteomic study which analyzed a hyperLOPIT data set from pluripotent mouse embryonic stem cells and predicted protein subcellular localization based on SVM.[160] As a result, the combination of multiple experimental data sets significantly improved the accuracy of protein subcellular assignment. Besides, several studies have combined *Arabidopsis* data sets from LOPIT experiments performed on different separation gradients to improve the organelle resolution and increase the coverage of the proteome.[137,138] Furthermore, transfer learning with auxiliary data can also improve the generalization accuracy of protein localization prediction. A transfer learning framework integrating heterogeneous data sources has been developed based on kNN and SVM, and its effectiveness has been validated in five LOPIT data sets from four different species.[43]

A variety of analytical tools have been developed to systematically analyze MS-based spatial proteomic data and achieve the above application purposes. The MSnbase is an R package for MS data processing (such as kNN missing data imputation), quality control and visualization.[188] The pRoloc package provides a complete data analysis framework for quantitative MS-based spatial proteomic data. It integrates various unsupervised (such as PCA) and supervised (such as RF, SVM, NB and neural networks) ML methods for data exploration, including data visualization, quality control and classification. The semi-supervised *phenoDisco* algorithm for novelty detection has also been integrated in the pRoloc package.[110] The pRolocGUI package is developed based on pRoloc, and it can be applied to visualize spatial proteomic data by PCA and display the prediction results from supervised ML such as SVM.[107] These packages are all available in the open-source Bioconductor project.[195] Other tools are also available to predict protein subcellular localization, such as MetaMass. MetaMass tool uses K-means clustering to identify protein clusters based on the marker set and then automatically maps the subcellular localization of unannotated proteins.[108]

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

The robustness of MetaMass has been validated in 11 subcellular proteomic data sets with expected results. There are also several tools for predicting changes in protein localization. TRANSPIRE tool is a dedicated analytical framework for the prediction of protein translocation. It performs predictions by training a stochastic variational Gaussian process classifier. The training data are synthesized by concatenation the spatial profiles of markers from two different organelles. TRANSPIRE has been applied to reveal protein movements during virus infections.[189]

4.2. Imaging-Based Spatial Proteomic Applications. The main purpose of imaging-based spatial proteomic data analysis is to identify the major subcellular patterns in cell images based on image features. Various supervised ML methods have been utilized to make this prediction, such as logistic regression, SVM and RF.[153,196] To determine the subcellular localization of new proteins in the test dataset, a prediction pipeline integrating K-means clustering and SVM classifier was developed to generalize across different proteins based on local features.[151] Ensemble learning is also a useful strategy to improve the predictive accuracy of image classification. The ensLOC is a typical ensemble model of SVM classifiers, which can assign each yeast cell image to subcellular localization classes on the basis of morphological image features.[46] The final result for each protein is 16 quantitative localization scores that define the proportion of protein-associated cell images assigned to each of 16 compartments. The ensLOC model has been applied to detect changes in yeast protein abundance and localization under environmental and genetic perturbations.[122] In addition, the subcellular distribution pattern of proteins can be used to train a ML to automatically distinguish between ferroptosis and apoptosis cells.[197] The investigation of organelle proteome heterogeneity in neighboring cells can reveal cell interactions and disease-related microenvironment features in diseases, such as systemic autoimmune disease[114] and breast cancer.[198,199]

Cell segmentation is an important step in imaging-based spatial proteomic data analysis and application, and various advanced CNN-based cell segmentation methods have been developed. RAMCES method uses a pretrained CNN to rank the cell boundary markers and combines top markers to construct weighted images for subsequent membrane-based cell segmentation methods.[139] The weighted images have proved successful in improving the accuracy of cell segmentation. FeatureNet is also a CNN model which can perform nuclear and cytoplasmic segmentation, and it has been used to demonstrate the spatial arrangement of immune-related proteins at subcellular resolution and reveal the tumor-immune microenvironment in triple negative breast cancer.[140,186] To ensure both accuracy and speed of cell segmentation, Mesmer, a CNN-based approach, has been created by integrating nuclear and whole-cell segmentation. The input of Mesmer consists of a nuclear image and a cytoplasmic image. It can automatically extract the characteristics of protein subcellular localization and has been used to detect the changes in cell morphology during human gestation.[156] Besides, semi-supervised AE is also effective in learning cell shapes. For example, MIRIAM is a robust pipeline for cell segmentation and shape characterization, which uses a convolutional AE to characterize cell shape.[174] The t-SNE analysis of latent representation obtained by MIRIAM showed that cells with similar shape clustered together, which indicated the power of AE in cell shape learning.

To fully leverage the strength of DL in imaging-based spatial proteomic data analysis, several strategies have been developed to accurately identify the protein localizations in images. Transfer learning also works in this process. An image features based multi-label DNN model, named as Loc-CAT, was built using image data from HPA Cell Atlas and showed high level of accuracy in multi-label prediction.[166] When using transfer learning to combine both manually annotated features and computational features, the accuracy of Loc-CAT in predicting protein subcellular localization was further enhanced. In addition, using pretrained DL models is also a robust transfer learning approach. For example, after pretraining the CNN-based DeepYeast on training dataset, the pretrained DeepYeast was used to extract features for images from another dataset containing new classes. The RF classifier using features learned by pretrained DeepYeast outperformed those using image features calculated by CellProfiler, indicating the generality of pretrained DeepYeast in extracting spatial patterns from yeast images.[146] Ensemble learning, multiple instance learning and self-supervised learning have also proved successful in identification of protein subcellular localization patterns.[155,159,183] Cytoself is an advanced self-supervised approach based on the vector quantized variational AE, and it can reduce images to latent representations to capture protein localization features.[171] It has been applied to analyze images from OpenCell. The representation of a specific protein was obtained by averaging latent representations from all the protein-associated images. UMAP was then used to visualize these protein encodings, and this analysis revealed the unique property of RNA binding proteins.[92]

Many imaging-based spatial proteomic data analysis tools provide ML-based analytical modules. ImageJ and Ilastik are two popular open-source software for image analysis, and they can be used for cell segmentation and image classification based on ML methods.[141,142] Some dedicated tools for cell segmentation are available. For instance, Cellpose is a tool for whole-cell segmentation which does not require any model retraining. It was developed on the basis of the U-Net architecture and was suitable for various types of cell images.[172] Squidpy provides numerous analysis methods that allow scalable analysis of spatial proteomic data. It integrates cell segmentation methods (such as Cellpose) and DL-based features extraction methods.[190] Cytokit and CellProfiler can also be used for image feature extraction and cell segmentation, and both tools provide advanced DL methods for cell image analysis.[157,191] As a part of CellProfiler, CellProfiler Analyst is a user-friendly tool that provides multiple ML algorithms for the construction of classifiers identifying various cell phenotypes.[200] To analyze large data set, the DeepCell Kiosk, as a cloud-native software, provides pretrained DL models for spatial proteomic image analysis, such as Mesmer.[111] There are also some tools which provide data visualization methods for spatial proteomic images. Giotto is a comprehensive toolbox which integrates various unsupervised dimensionality reduction methods (such as PCA, UMAP and t-SNE) and clustering methods (such as K-means and hierarchical clustering). It can be applied to detect subcellular protein localization patterns.[192] Cellar is an open-source tool which also supports these unsupervised ML methods for image data visualization, and it has been used to analyze a lymph node CODEX spatial

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

proteomic data set.[115] These analytical tools lay a solid foundation for imaging-based spatial proteomics.

5. CHALLENGES AND FUTURE DIRECTIONS

Recent advances in ML technology have facilitated the development of spatial proteomics. However, there are still several computational challenges in ML-aided spatial proteomic applications. The lack of systematic data resources is a fundamental problem. Although many data resources are currently available, they only provide spatial proteomic data in specific species, cell types or conditions.[3] To ensure the accessibility and reusability of these valuable data, it is of great importance to construct an overarching and dedicated spatial proteomic data repository that provides raw data, processed quantitative data and qualitative data genrated by different methodologies (both MS-based and imaging-based approches).[1,21] In addition, the reliability of organelle markers can significantly affect the accuracy of protein localization prediction. In addition, the reliability of organelle markers can significantly affect the accuracy of protein localization prediction. Most of the currently available markers are annotated in a default state, rather than in specific research systems or under specific conditions.[26] Due to the absence of such detailed annotations, some studies did not use strictly defined markers. In order to provide reliable markers under specific conditions, some preliminary attempts have been made. For example, the Translocatome database provides protein subcellular localization information under different conditions, but it is limited to a few human proteins.[101] More efforts should be made to identify reliable organelle markers and complete their annotations.

Multimodal data integration is also a challenge in spatial proteomics, including the combination of spatially resolved and temporally resolved proteomic data and the integrative analysis of multiomics data. The major obstacles in integrating data from multiple modalities are the existence of batch effects and the distinct feature spaces of different omics data.[22,201] Integrating MS-based and imaging-based spatial proteomics approaches allows simultaneous analyses of multiple organelles at different time points, which can help understand disease processes. Advanced data analysis pipelines for such spatiotemporal analyses are desired to better integrate spatial and temporal maps of proteome changes during disease progression.[21,23] Since the occurrence of disease changes multiple interacting components of biological systems (such as RNA, proteins, lipids, and metabolites), multiomics integration incorporating spatial proteomics can give unprecedented insights into cell functionality.[2,23] Combining spatial proteomics and other omics data enables comprehensive studies of the heterogeneity in various diseases, and numerous related investigations have been conducted.[8,202−204] The development of analytical tools that provide workflows for multimodal data integration is currently a big challenge.[202] Moreover, the lack of unified and systematic data analysis tools hampers the comparative analyses of changes in protein spatial patterns under different conditions.[22] The development and application of ML models which are designed to take multiple conditions into consideration can facilitate the exploration of the subcellular proteomic dynamics.[21]

The interpretability of ML models is one of the biggest challenges in ML-aided spatial proteomic applications. Most ML methods applied in current spatial proteomics lack meaningful biological assumptions.[22] The development of interpretable ML models provides an opportunity to analyze the identified spatial patterns from a biological perspective.[1]

6. CONCLUSION

The remarkable advances of spatial proteomics provide powerful tools for in-depth exploration of cell biology. A large number of spatial proteomic data repositories are currently available and lay a solid foundation for the application of ML methods in spatial proteomic data analysis. Various ML methods have been successfully applied, and different types of ML algorithms play different roles in the analytical workflows. Although a variety of analytical tools integrating state-of-the-art ML methods have been developed for diverse spatial proteomic applications, it is certain that more efforts should be made to develop overarching, systematic and interpretable methods to facilitate the reliable practice of biomedical research. In summary, the computational advances discussed in this review provide valuable guidelines for cell biologists who will engage in the spatial proteomic research. ML-aided spatial proteomic data analysis paves the way to unravel cell biology and inspire the medical and drug discovery communities.

■ ASSOCIATED CONTENT

Data Availability Statement

The URLs of a variety of data resources and analytical tools for spatial proteomics are available in Tables 1 and 2.

■ AUTHOR INFORMATION

Corresponding Author

Feng Zhu − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0001-8069-0053; Email: zhufeng@zju.edu.cn

Authors

Minjie Mou − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0001-7619-2975

Ziqi Pan − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-3883-4161

Mingkun Lu − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0003-1522-6320

Huaicheng Sun − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-1381-9571

Yunxia Wang − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0003-1951-942X

Yongchao Luo − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-4140-5392

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01161

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by Natural Science Foundation of Zhejiang Province (LR21H300001), National Natural Science

Foundation of China (U1909208 and 81872798), 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 and Biomedicine), Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Alibaba Cloud, Information Tech Center of Zhejiang University.

■ ABBREVIATIONS

AE, autoencoder; AP-MS, affinity purification-mass spectrometry; BBBC, Broad Bioimage Benchmark Collection; BioID, biotin identification; CIL, Cell Image Library; CNN, convolutional neural network; CODEX, CO-Detection by indEXing; CYCLoPs, Collection of Yeast Cells Localization Patterns; dHITS, Database of High Throughput Screening Hits; DL, deep learning; DNN, deep neural network; eSLDB, eukaryotic Subcellular Localization DataBase; GFP, green fluorescent protein; GMM, Gaussian mixture model; GO, Gene Ontology; HCMC, human cytomegalovirus; HEK, human embryonic kidney; HPA, Human Protein Atlas; HuBMAP, Human Biomolecular Atlas Program; IDR, Image Data Resource; kNN, k-nearest neighbors; LOPIT, localization of organelle proteins by isotope tagging; LoQAtE, Localization and Quantitation Atlas of the yeast proteomE database; MassIVE, Mass Spectrometry Interactive Virtual Environment; ML, machine learning; MS, mass spectrometry; NAFLD, nonalcoholic fatty liver disease; NB, naive Bayesian; NMF, nonnegative matrix factorization; PCA, principal component analysis; PCP, protein correlation profiling; PLS-DA, partial least-squares discriminant analysis; PPI, protein−protein interactions; PRIDE, PRoteomics IDEntifications Database; RF, random forest; SUBA4, SUBcellular location database for *Arabidopsis* proteins; SVM, support vector machine; SVM-RFE, support-vector-machine-based recursive-feature elimination; t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection; UniProt, Universal Protein Resource; YeastGFP, Yeast GFP Fusion Localization Database; YPL+, Yeast Protein Localization Plus Database; YRC PIR, Yeast Resource Center Public Image Repository

■ REFERENCES

(1) Lundberg, E.; Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 285−302.

(2) Palla, G.; Fischer, D. S.; Regev, A.; Theis, F. J. Spatial components of molecular tissue biology. *Nat. Biotechnol.* **2022**, *40*, 308−318.

(3) Singh, A. Subcellular proteome map of human cells. *Nat. Methods.* **2021**, *18*, 713.

(4) Watson, J.; Smith, M.; Francavilla, C.; Schwartz, J. M. SubcellulaRVis: a web-based tool to simplify and visualise subcellular compartment enrichment. *Nucleic Acids Res.* **2022**, *50*, W718−W725.

(5) Thul, P. J.; Akesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Ait Blal, H.; Alm, T.; Asplund, A.; Bjork, L.; Breckels, L. M.; Backstrom, A.; Danielsson, F.; Fagerberg, L.; Fall, J.; Gatto, L.; Gnann, C.; Hober, S.; Hjelmare, M.; Johansson, F.; Lee, S.; Lindskog, C.; Mulder, J.; Mulvey, C. M.; Nilsson, P.; Oksvold, P.; Rockberg, J.; Schutten, R.; Schwenk, J. M.; Sivertsson, A.; Sjostedt, E.; Skogs, M.; Stadler, C.; Sullivan, D. P.; Tegel, H.; Winsnes, C.; Zhang, C.; Zwahlen, M.; Mardinoglu, A.; Ponten, F.; von Feilitzen, K.; Lilley, K. S.; Uhlen, M.; Lundberg, E. A subcellular map of the human proteome. *Science.* **2017**, *356*, No. eaal3321.

(6) Niemi, N. M.; MacKeigan, J. P. Mitochondrial phosphorylation in apoptosis: flipping the death switch. *Antioxid Redox Signal.* **2013**, *19*, 572−582.

(7) He, Y.; Yu, Z.; Ge, D.; Wang-Sattler, R.; Thiesen, H. J.; Xie, L.; Li, Y. Cell type specificity of signaling: view from membrane receptors distribution and their downstream transduction networks. *Protein Cell.* **2012**, *3*, 701−713.

(8) Mahdessian, D.; Cesnik, A. J.; Gnann, C.; Danielsson, F.; Stenstrom, L.; Arif, M.; Zhang, C.; Le, T.; Johansson, F.; Schutten, R.; Backstrom, A.; Axelsson, U.; Thul, P.; Cho, N. H.; Carja, O.; Uhlen, M.; Mardinoglu, A.; Stadler, C.; Lindskog, C.; Ayoglu, B.; Leonetti, M. D.; Ponten, F.; Sullivan, D. P.; Lundberg, E. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature.* **2021**, *590*, 649−654.

(9) Shin, S. J.; Smith, J. A.; Rezniczek, G. A.; Pan, S.; Chen, R.; Brentnall, T. A.; Wiche, G.; Kelly, K. A. Unexpected gain of function for the scaffolding protein plectin due to mislocalization in pancreatic cancer. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 19414−19419.

(10) Neel, D. S.; Allegakoen, D. V.; Olivas, V.; Mayekar, M. K.; Hemmati, G.; Chatterjee, N.; Blakely, C. M.; McCoach, C. E.; Rotow, J. K.; Le, A.; Karachaliou, N.; Rosell, R.; Riess, J. W.; Nichols, R.; Doebele, R. C.; Bivona, T. G. Differential Subcellular Localization Regulates Oncogenic Signaling by ROS1 Kinase Fusion Proteins. *Cancer Res.* **2019**, *79*, 546−556.

(11) Guardia, C. M.; De Pace, R.; Mattera, R.; Bonifacino, J. S. Neuronal functions of adaptor complexes involved in protein sorting. *Curr. Opin Neurobiol.* **2018**, *51*, 103−110.

(12) Meyer, K.; Kirchner, M.; Uyar, B.; Cheng, J. Y.; Russo, G.; Hernandez-Miranda, L. R.; Szymborska, A.; Zauber, H.; Rudolph, I. M.; Willnow, T. E.; Akalin, A.; Haucke, V.; Gerhardt, H.; Birchmeier, C.; Kuhn, R.; Krauss, M.; Diecke, S.; Pascual, J. M.; Selbach, M. Mutations in Disordered Regions Can Cause Disease by Creating Dileucine Motifs. *Cell.* **2018**, *175*, 239−253.

(13) O'Neill, A. C.; Uzbas, F.; Antognolli, G.; Merino, F.; Draganova, K.; Jack, A.; Zhang, S.; Pedini, G.; Schessner, J. P.; Cramer, K.; Schepers, A.; Metzger, F.; Esgleas, M.; Smialowski, P.; Guerrini, R.; Falk, S.; Feederle, R.; Freytag, S.; Wang, Z.; Bahlo, M.; Jungmann, R.; Bagni, C.; Borner, G. H. H.; Robertson, S. P.; Hauck, S. M.; Gotz, M. Spatial centrosome proteome of human neural cells uncovers disease-relevant heterogeneity. *Science.* **2022**, *376*, No. eabf9088.

(14) Liao, Y. C.; Fernandopulle, M. S.; Wang, G.; Choi, H.; Hao, L.; Drerup, C. M.; Patel, R.; Qamar, S.; Nixon-Abell, J.; Shen, Y.; Meadows, W.; Vendruscolo, M.; Knowles, T. P. J.; Nelson, M.; Czekalska, M. A.; Musteikyte, G.; Gachechiladze, M. A.; Stephens, C. A.; Pasolli, H. A.; Forrest, L. R.; St George-Hyslop, P.; Lippincott-Schwartz, J.; Ward, M. E. RNA Granules Hitchhike on Lysosomes for Long-Distance Transport, Using Annexin A11 as a Molecular Tether. *Cell.* **2019**, *179*, 147−164.

(15) Siljee, J. E.; Wang, Y.; Bernard, A. A.; Ersoy, B. A.; Zhang, S.; Marley, A.; Von Zastrow, M.; Reiter, J. F.; Vaisse, C. Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. *Nat. Genet.* **2018**, *50*, 180−185.

(16) Krahmer, N.; Najafi, B.; Schueder, F.; Quagliarini, F.; Steger, M.; Seitz, S.; Kasper, R.; Salinas, F.; Cox, J.; Uhlenhaut, N. H.; Walther, T. C.; Jungmann, R.; Zeigerer, A.; Borner, G. H. H.; Mann, M. Organellar Proteomics and Phospho-Proteomics Reveal Subcellular Reorganization in Diet-Induced Hepatic Steatosis. *Dev Cell.* **2018**, *47*, 205−221.

(17) Pankow, S.; Martinez-Bartolome, S.; Bamberger, C.; Yates, J. R. Understanding molecular mechanisms of disease through spatial proteomics. *Curr. Opin Chem. Biol.* **2019**, *48*, 19−25.

(18) Itzhak, D. N.; Davies, C.; Tyanova, S.; Mishra, A.; Williamson, J.; Antrobus, R.; Cox, J.; Weekes, M. P.; Borner, G. H. H. A Mass Spectrometry-Based Approach for Mapping Protein Subcellular

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

Localization Reveals the Spatial Proteome of Mouse Primary Neurons. *Cell Rep.* **2017**, *20*, 2706−2718.
(19) Kim, D. I.; Roux, K. J. Filling the Void: Proximity-Based Labeling of Proteins in Living Cells. *Trends Cell Biol.* **2016**, *26*, 804−817.
(20) Torres, N. P.; Ho, B.; Brown, G. W. High-throughput fluorescence microscopic analysis of protein abundance and localization in budding yeast. *Crit Rev. Biochem Mol. Biol.* **2016**, *51*, 110−119.
(21) Christopher, J. A.; Stadler, C.; Martin, C. E.; Morgenstern, M.; Pan, Y.; Betsinger, C. N.; Rattray, D. G.; Mahdessian, D.; Gingras, A. C.; Warscheid, B.; Lehtio, J.; Cristea, I. M.; Foster, L. J.; Emili, A.; Lilley, K. S. Subcellular proteomics. *Nat. Rev. Methods Primers.* **2021**, *22*, 27−56.
(22) Zhang, M.; Sheffield, T.; Zhan, X.; Li, Q.; Yang, D. M.; Wang, Y.; Wang, S.; Xie, Y.; Wang, T.; Xiao, G. Spatial molecular profiling: platforms, applications and analysis tools. *Brief Bioinform.* **2021**, *22*, bbaa145.
(23) Rahmatbakhsh, M.; Gagarinova, A.; Babu, M. Bioinformatic Analysis of Temporal and Spatial Proteome Alternations During Infections. *Front Genet.* **2021**, *12*, 667936.
(24) Du, P.; Li, T.; Wang, X. Recent progress in predicting protein sub-subcellular locations. *Expert Rev. Proteomics.* **2011**, *8*, 391−404.
(25) Huang, W. L.; Tung, C. W.; Ho, S. W.; Hwang, S. F.; Ho, S. Y. ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics.* **2008**, *9*, 80.
(26) Gatto, L.; Breckels, L. M.; Burger, T.; Nightingale, D. J.; Groen, A. J.; Campbell, C.; Nikolovski, N.; Mulvey, C. M.; Christoforou, A.; Ferro, M.; Lilley, K. S. A foundation for reliable spatial proteomics data analysis. *Mol. Cell Proteomics.* **2014**, *13*, 1937−1952.
(27) Tharkeshwar, A. K.; Gevaert, K.; Annaert, W. Organellar Omics-A Reviving Strategy to Untangle the Biomolecular Complexity of the Cell. *Proteomics.* **2018**, *18*, No. e1700113.
(28) Peikert, C. D.; Mani, J.; Morgenstern, M.; Kaser, S.; Knapp, B.; Wenger, C.; Harsman, A.; Oeljeklaus, S.; Schneider, A.; Warscheid, B. Charting organellar importomes by quantitative mass spectrometry. *Nat. Commun.* **2017**, *8*, 15272.
(29) Foster, L. J.; de Hoog, C. L.; Zhang, Y.; Zhang, Y.; Xie, X.; Mootha, V. K.; Mann, M. A mammalian organelle map by protein correlation profiling. *Cell.* **2006**, *125*, 187−199.
(30) Fagerberg, L.; Stromberg, S.; El-Obeid, A.; Gry, M.; Nilsson, K.; Uhlen, M.; Ponten, F.; Asplund, A. Large-scale protein profiling in human cell lines using antibody-based proteomics. *J. Proteome Res.* **2011**, *10*, 4066−4075.
(31) Dunkley, T. P.; Hester, S.; Shadforth, I. P.; Runions, J.; Weimar, T.; Hanton, S. L.; Griffin, J. L.; Bessant, C.; Brandizzi, F.; Hawes, C.; Watson, R. B.; Dupree, P.; Lilley, K. S. Mapping the Arabidopsis organelle proteome. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 6518−6523.
(32) Sadowski, P. G.; Dunkley, T. P.; Shadforth, I. P.; Dupree, P.; Bessant, C.; Griffin, J. L.; Lilley, K. S. Quantitative proteomic approach to study subcellular localization of membrane proteins. *Nat. Protoc.* **2006**, *1*, 1778−1789.
(33) Lu, A. X.; Chong, Y. T.; Hsu, I. S.; Strome, B.; Handfield, L. F.; Kraus, O.; Andrews, B. J.; Moses, A. M. Integrating images from multiple microscopy screens reveals diverse patterns of change in the subcellular localization of proteins. *Elife.* **2018**, *7*, No. e31872.
(34) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noe, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121*, 9722−9758.
(35) Caragea, C.; Caragea, D.; Silvescu, A.; Honavar, V. Semi-supervised prediction of protein subcellular localization using abstraction augmented Markov models. *BMC Bioinformatics.* **2010**, *11*, S6.
(36) Xu, Q.; Hu, D. H.; Xue, H.; Yu, W.; Yang, Q. Semi-supervised protein subcellular localization. *BMC Bioinformatics.* **2009**, *10*, S47.
(37) Ohta, S.; Bukowski-Wills, J. C.; Sanchez-Pulido, L.; de Lima Alves, F.; Wood, L.; Chen, Z. A.; Platani, M.; Fischer, L.; Hudson, D. F.; Ponting, C. P.; Fukagawa, T.; Earnshaw, W. C.; Rappsilber, J. The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell.* **2010**, *142*, 810−821.
(38) Swan, A. L.; Mobasheri, A.; Allaway, D.; Liddell, S.; Bacardit, J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS.* **2013**, *17*, 595−610.
(39) Schnell, U.; Dijk, F.; Sjollema, K. A.; Giepmans, B. N. Immunolabeling artifacts and the need for live-cell imaging. *Nat. Methods.* **2012**, *9*, 152−158.
(40) Kraus, O. Z.; Grys, B. T.; Ba, J.; Chong, Y.; Frey, B. J.; Boone, C.; Andrews, B. J. Automated analysis of high-content microscopy data with deep learning. *Mol. Syst. Biol.* **2017**, *13*, 924.
(41) Hu, H.; Bindu, J. P.; Laskin, J. Self-supervised clustering of mass spectrometry imaging data using contrastive learning. *Chem. Sci.* **2021**, *13*, 90−98.
(42) Wang, Y.; Magar, R.; Liang, C.; Barati Farimani, A. Improving Molecular Contrastive Learning via Faulty Negative Mitigation and Decomposed Fragment Contrast. *J. Chem. Inf Model.* **2022**, *62*, 2713−2725.
(43) Breckels, L. M.; Holden, S. B.; Wojnar, D.; Mulvey, C. M.; Christoforou, A.; Groen, A.; Trotter, M. W.; Kohlbacher, O.; Lilley, K. S.; Gatto, L. Learning from Heterogeneous Data Sources: An Application in Spatial Proteomics. *PLoS Comput. Biol.* **2016**, *12*, No. e1004920.
(44) Cheplygina, V.; de Bruijne, M.; Pluim, J. P. W. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **2019**, *54*, 280−296.
(45) Karimi, D.; Warfield, S. K.; Gholipour, A. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. *Artif Intell Med.* **2021**, *116*, 102078.
(46) Koh, J. L.; Chong, Y. T.; Friesen, H.; Moses, A.; Boone, C.; Andrews, B. J.; Moffat, J. CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in Saccharomyces cerevisiae. *G3 (Bethesda).* **2015**, *5*, 1223−1232.
(47) Li, J.; Zhang, L.; He, S.; Guo, F.; Zou, Q. SubLocEP: a novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning. *Brief Bioinform.* **2021**, *22*, bbaa401.
(48) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods.* **2016**, *13*, 731−740.
(49) Caicedo, J. C.; Cooper, S.; Heigwer, F.; Warchal, S.; Qiu, P.; Molnar, C.; Vasilevich, A. S.; Barry, J. D.; Bansal, H. S.; Kraus, O.; Wawer, M.; Paavolainen, L.; Herrmann, M. D.; Rohban, M.; Hung, J.; Hennig, H.; Concannon, J.; Smith, I.; Clemons, P. A.; Singh, S.; Rees, P.; Horvath, P.; Linington, R. G.; Carpenter, A. E. Data-analysis strategies for image-based cell profiling. *Nat. Methods.* **2017**, *14*, 849−863.
(50) Andersen, J. S.; Wilkinson, C. J.; Mayor, T.; Mortensen, P.; Nigg, E. A.; Mann, M. Proteomic characterization of the human centrosome by protein correlation profiling. *Nature.* **2003**, *426*, 570−574.
(51) Bouchnak, I.; Brugiere, S.; Moyet, L.; Le Gall, S.; Salvi, D.; Kuntz, M.; Tardif, M.; Rolland, N. Unraveling Hidden Components of the Chloroplast Envelope Proteome: Opportunities and Limits of Better MS Sensitivity. *Mol. Cell Proteomics.* **2019**, *18*, 1285−1306.
(52) Krahmer, N.; Hilger, M.; Kory, N.; Wilfling, F.; Stoehr, G.; Mann, M.; Farese, R. V., Jr.; Walther, T. C. Protein correlation profiles identify lipid droplet proteins with high confidence. *Mol. Cell Proteomics.* **2013**, *12*, 1115−1126.
(53) Dunkley, T. P.; Watson, R.; Griffin, J. L.; Dupree, P.; Lilley, K. S. Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell Proteomics.* **2004**, *3*, 1128−1134.
(54) Geladaki, A.; Kocevar Britovsek, N.; Breckels, L. M.; Smith, T. S.; Vennard, O. L.; Mulvey, C. M.; Crook, O. M.; Gatto, L.; Lilley, K.

S. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* **2019**, *10*, 331.
(55) Itzhak, D. N.; Tyanova, S.; Cox, J.; Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife.* **2016**, *5*, No. e16950.
(56) Kozik, P.; Gros, M.; Itzhak, D. N.; Joannas, L.; Heurtebise-Chretien, S.; Krawczyk, P. A.; Rodriguez-Silvestre, P.; Alloatti, A.; Magalhaes, J. G.; Del Nery, E.; Borner, G. H. H.; Amigorena, S. Small Molecule Enhancers of Endosome-to-Cytosol Import Augment Anti-tumor Immunity. *Cell Rep.* **2020**, *32*, 107905.
(57) De Duve, C.; Pressman, B. C.; Gianetto, R.; Wattiaux, R.; Appelmans, F. Tissue fractionation studies. 6. Intracellular distribution patterns of enzymes in rat-liver tissue. *Biochem. J.* **1955**, *60*, 604−617.
(58) Huttlin, E. L.; Bruckner, R. J.; Paulo, J. A.; Cannon, J. R.; Ting, L.; Baltier, K.; Colby, G.; Gebreab, F.; Gygi, M. P.; Parzen, H.; Szpyt, J.; Tam, S.; Zarraga, G.; Pontano-Vaites, L.; Swarup, S.; White, A. E.; Schweppe, D. K.; Rad, R.; Erickson, B. K.; Obar, R. A.; Guruharsha, K. G.; Li, K.; Artavanis-Tsakonas, S.; Gygi, S. P.; Harper, J. W. Architecture of the human interactome defines protein communities and disease networks. *Nature.* **2017**, *545*, 505−509.
(59) Lonn, P.; Landegren, U. Close Encounters - Probing Proximal Proteins in Live or Fixed Cells. *Trends Biochem. Sci.* **2017**, *42*, 504−515.
(60) Rhee, H. W.; Zou, P.; Udeshi, N. D.; Martell, J. D.; Mootha, V. K.; Carr, S. A.; Ting, A. Y. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science.* **2013**, *339*, 1328−1331.
(61) Kim, D. I.; Birendra, K. C.; Zhu, W.; Motamedchaboki, K.; Doye, V.; Roux, K. J. Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, No. e2453-E2461.
(62) Roux, K. J.; Kim, D. I.; Raida, M.; Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **2012**, *196*, 801−810.
(63) Tang, Y.; Huang, A.; Gu, Y. Global profiling of plant nuclear membrane proteome in Arabidopsis. *Nat. Plants.* **2020**, *6*, 838−847.
(64) Stadler, C.; Skogs, M.; Brismar, H.; Uhlen, M.; Lundberg, E. A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics.* **2010**, *73*, 1067−1078.
(65) Stadler, C.; Rexhepaj, E.; Singan, V. R.; Murphy, R. F.; Pepperkok, R.; Uhlen, M.; Simpson, J. C.; Lundberg, E. Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat. Methods.* **2013**, *10*, 315−323.
(66) Tiede, C.; Bedford, R.; Heseltine, S. J.; Smith, G.; Wijetunga, I.; Ross, R.; AlQallaf, D.; Roberts, A. P.; Balls, A.; Curd, A.; Hughes, R. E.; Martin, H.; Needham, S. R.; Zanetti-Domingues, L. C.; Sadigh, Y.; Peacock, T. P.; Tang, A. A.; Gibson, N.; Kyle, H.; Platt, G. W.; Ingram, N.; Taylor, T.; Coletta, L. P.; Manfield, I.; Knowles, M.; Bell, S.; Esteves, F.; Maqbool, A.; Prasad, R. K.; Drinkhill, M.; Bon, R. S.; Patel, V.; Goodchild, S. A.; Martin-Fernandez, M.; Owens, R. J.; Nettleship, J. E.; Webb, M. E.; Harrison, M.; Lippiat, J. D.; Ponnambalam, S.; Peckham, M.; Smith, A.; Ferrigno, P. K.; Johnson, M.; McPherson, M. J.; Tomlinson, D. C. Affimer proteins are versatile and renewable affinity reagents. *Elife.* **2017**, *6*, No. e24903.
(67) Marx, V. Calling the next generation of affinity reagents. *Nat. Methods.* **2013**, *10*, 829−833.
(68) Denervaud, N.; Becker, J.; Delgado-Gonzalo, R.; Damay, P.; Rajkumar, A. S.; Unser, M.; Shore, D.; Naef, F.; Maerkl, S. J. A chemostat array enables the spatio-temporal analysis of the yeast proteome. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 15842−15847.
(69) Cong, L.; Ran, F. A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A.; Zhang, F. Multiplex genome engineering using CRISPR/Cas systems. *Science.* **2013**, *339*, 819−823.
(70) de Groot, R.; Luthi, J.; Lindsay, H.; Holtackers, R.; Pelkmans, L. Large-scale image-based profiling of single-cell phenotypes in arrayed CRISPR-Cas9 gene perturbation screens. *Mol. Syst. Biol.* **2018**, *14*, No. e8064.
(71) Mattiazzi Usaj, M.; Styles, E. B.; Verster, A. J.; Friesen, H.; Boone, C.; Andrews, B. J. High-Content Screening for Quantitative Cell Biology. *Trends Cell Biol.* **2016**, *26*, 598−611.
(72) Orre, L. M.; Vesterlund, M.; Pan, Y.; Arslan, T.; Zhu, Y.; Fernandez Woodbridge, A.; Frings, O.; Fredlund, E.; Lehtio, J. SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol. Cell* **2019**, *73*, 166−182.
(73) Huang, F.; Tang, X.; Ye, B.; Wu, S.; Ding, K. PSL-LCCL: a resource for subcellular protein localization in liver cancer cell line SK_HEP1. *Database (Oxford).* **2022**, *2022*, baab087.
(74) Go, C. D.; Knight, J. D. R.; Rajasekharan, A.; Rathod, B.; Hesketh, G. G.; Abe, K. T.; Youn, J. Y.; Samavarchi-Tehrani, P.; Zhang, H.; Zhu, L. Y.; Popiel, E.; Lambert, J. P.; Coyaud, E.; Cheung, S. W. T.; Rajendran, D.; Wong, C. J.; Antonicka, H.; Pelletier, L.; Palazzo, A. F.; Shoubridge, E. A.; Raught, B.; Gingras, A. C. A proximity-dependent biotinylation map of a human cell. *Nature.* **2021**, *595*, 120−124.
(75) Huttlin, E. L.; Bruckner, R. J.; Navarrete-Perea, J.; Cannon, J. R.; Baltier, K.; Gebreab, F.; Gygi, M. P.; Thornock, A.; Zarraga, G.; Tam, S.; Szpyt, J.; Gassaway, B. M.; Panov, A.; Parzen, H.; Fu, S.; Golbazi, A.; Maenpaa, E.; Stricker, K.; Guha Thakurta, S.; Zhang, T.; Rad, R.; Pan, J.; Nusinow, D. P.; Paulo, J. A.; Schweppe, D. K.; Vaites, L. P.; Harper, J. W.; Gygi, S. P. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell.* **2021**, *184*, 3022−3040.
(76) Christoforou, A.; Mulvey, C. M.; Breckels, L. M.; Geladaki, A.; Hurrell, T.; Hayward, P. C.; Naake, T.; Gatto, L.; Viner, R.; Martinez Arias, A.; Lilley, K. S. A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* **2016**, *7*, 8992.
(77) Jadot, M.; Boonen, M.; Thirion, J.; Wang, N.; Xing, J.; Zhao, C.; Tannous, A.; Qian, M.; Zheng, H.; Everett, J. K.; Moore, D. F.; Sleat, D. E.; Lobel, P. Accounting for Protein Subcellular Localization: A Compartmental Map of the Rat Liver Proteome. *Mol. Cell Proteomics.* **2017**, *16*, 194−212.
(78) Perez-Riverol, Y.; Bai, J.; Bandla, C.; Garcia-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaino, J. A. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **2022**, *50*, D543−D552.
(79) Sharma, V.; Eckels, J.; Taylor, G. K.; Shulman, N. J.; Stergachis, A. B.; Joyner, S. A.; Yan, P.; Whiteaker, J. R.; Halusa, G. N.; Schilling, B.; Gibson, B. W.; Colangelo, C. M.; Paulovich, A. G.; Carr, S. A.; Jaffe, J. D.; MacCoss, M. J.; MacLean, B. Panorama: a targeted proteomics knowledge base. *J. Proteome Res.* **2014**, *13*, 4205−4210.
(80) Deutsch, E. W.; Lam, H.; Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **2008**, *9*, 429−434.
(81) Wang, M.; Wang, J.; Carver, J.; Pullman, B. S.; Cha, S. W.; Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst.* **2018**, *7*, 412−421.
(82) Huh, W. K.; Falvo, J. V.; Gerke, L. C.; Carroll, A. S.; Howson, R. W.; Weissman, J. S.; O'Shea, E. K. Global analysis of protein localization in budding yeast. *Nature.* **2003**, *425*, 686−691.
(83) Breker, M.; Gymrek, M.; Moldavski, O.; Schuldiner, M. LoQAtE-Localization and Quantitation ATlas of the yeast proteomE.article-title > A new tool for multiparametric dissection of single-protein behavior in response to biological perturbations in yeast. *Nucleic Acids Res.* **2014**, *42*, D726−D730.
(84) Dubreuil, B.; Sass, E.; Nadav, Y.; Heidenreich, M.; Georgeson, J. M.; Weill, U.; Duan, Y.; Meurer, M.; Schuldiner, M.; Knop, M.; Levy, E. D. YeastRGB: comparing the abundance and localization of yeast proteins across cells and libraries. *Nucleic Acids Res.* **2019**, *47*, D1245−D1249.
(85) Chuartzman, S. G.; Schuldiner, M. Database for High Throughput Screening Hits (dHITS): a simple tool to retrieve gene

specific phenotypes from systematic screens done in yeast. *Yeast.* **2018**, *35*, 477−483.
(86) Riffle, M.; Davis, T. N. The Yeast Resource Center Public Image Repository: A large database of fluorescence microscopy images. *BMC Bioinformatics.* **2010**, *11*, 263.
(87) Kals, M.; Natter, K.; Thallinger, G. G.; Trajanoski, Z.; Kohlwein, S. D. YPL.db2: the Yeast Protein Localization database, version 2.0. *Yeast.* **2005**, *22*, 213−218.
(88) Frenkel-Morgenstern, M.; Cohen, A. A.; Geva-Zatorsky, N.; Eden, E.; Prilusky, J.; Issaeva, I.; Sigal, A.; Cohen-Saidon, C.; Liron, Y.; Cohen, L.; Danon, T.; Perzov, N.; Alon, U. Dynamic Proteomics: a database for dynamics and localizations of endogenous fluorescently-tagged proteins in living human cells. *Nucleic Acids Res.* **2010**, *38*, D508−D512.
(89) Orloff, D. N.; Iwasa, J. H.; Martone, M. E.; Ellisman, M. H.; Kane, C. M. The cell: an image library-CCDB: a curated repository of microscopy data. *Nucleic Acids Res.* **2012**, *41*, D1241−D1250.
(90) Williams, E.; Moore, J.; Li, S. W.; Rustici, G.; Tarkowska, A.; Chessel, A.; Leo, S.; Antal, B.; Ferguson, R. K.; Sarkans, U.; Brazma, A.; Salas, R. E. C.; Swedlow, J. R. The Image Data Resource: A Bioimage Data Integration and Publication Platform. *Nat. Methods.* **2017**, *14*, 775−781.
(91) Ljosa, V.; Sokolnicki, K. L.; Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nat. Methods.* **2012**, *9*, 637.
(92) Cho, N. H.; Cheveralls, K. C.; Brunner, A. D.; Kim, K.; Michaelis, A. C.; Raghavan, P.; Kobayashi, H.; Savy, L.; Li, J. Y.; Canaj, H.; Kim, J. Y. S.; Stewart, E. M.; Gnann, C.; McCarthy, F.; Cabrera, J. P.; Brunetti, R. M.; Chhun, B. B.; Dingle, G.; Hein, M. Y.; Huang, B.; Mehta, S. B.; Weissman, J. S.; Gomez-Sjoberg, R.; Itzhak, D. N.; Royer, L. A.; Mann, M.; Leonetti, M. D. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science.* **2022**, *375*, No. eabi6983.
(93) Bateman, A.; et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.
(94) The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **2019**, *47*, D330−D338.
(95) Pierleoni, A.; Martelli, P. L.; Fariselli, P.; Casadio, R. eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Res.* **2007**, *35*, D208−D212.
(96) Binder, J. X.; Pletscher-Frankild, S.; Tsafou, K.; Stolte, C.; O'Donoghue, S. I.; Schneider, R.; Jensen, L. J. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford).* **2014**, *2014*, bau012.
(97) Rastogi, S.; Rost, B. LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana. *Nucleic Acids Res.* **2011**, *39*, D230−D234.
(98) Veres, D. V.; Gyurko, D. M.; Thaler, B.; Szalay, K. Z.; Fazekas, D.; Korcsmaros, T.; Csermely, P. ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Res.* **2015**, *43*, D485−D493.
(99) Zhu, L.; Malatras, A.; Thorley, M.; Aghoghogbe, I.; Mer, A.; Duguez, S.; Butler-Browne, G.; Voit, T.; Duddy, W. CellWhere: graphical display of interaction networks organized on subcellular localizations. *Nucleic Acids Res.* **2015**, *43*, W571−W575.
(100) Hooper, C. M.; Castleden, I. R.; Tanz, S. K.; Aryamanesh, N.; Millar, A. H. SUBA4: the interactive data analysis centre for Arabidopsis subcellular protein locations. *Nucleic Acids Res.* **2017**, *45*, D1064−D1074.
(101) Mendik, P.; Dobronyi, L.; Hari, F.; Kerepesi, C.; Maia-Moco, L.; Buszlai, D.; Csermely, P.; Veres, D. V. Translocatome: a novel resource for the analysis of protein translocation between cellular organelles. *Nucleic Acids Res.* **2019**, *47*, D495−D505.
(102) Schluter, A.; Real-Chicharro, A.; Gabaldon, T.; Sanchez-Jimenez, F.; Pujol, A. PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. *Nucleic Acids Res.* **2010**, *38*, D800−D805.
(103) Rath, S.; Sharma, R.; Gupta, R.; Ast, T.; Chan, C.; Durham, T. J.; Goodman, R. P.; Grabarek, Z.; Haas, M. E.; Hung, W. H. W.; Joshi, P. R.; Jourdain, A. A.; Kim, S. H.; Kotrys, A. V.; Lam, S. S.; McCoy, J. G.; Meisel, J. D.; Miranda, M.; Panda, A.; Patgiri, A.; Rogers, R.; Sadre, S.; Shah, H.; Skinner, O. S.; To, T. L.; Walker, M. A.; Wang, H.; Ward, P. S.; Wengrod, J.; Yuan, C. C.; Calvo, S. E.; Mootha, V. K. MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.* **2021**, *49*, D1541−D1547.
(104) Smith, A. C.; Robinson, A. J. MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. *Nucleic Acids Res.* **2019**, *47*, D1225−D1228.
(105) Akhter, S.; Kaur, H.; Agrawal, P.; Raghava, G. P. S. RareLSD: a manually curated database of lysosomal enzymes associated with rare diseases. *Database (Oxford).* **2019**, *2019*, baz112.
(106) Barylyuk, K.; Koreny, L.; Ke, H.; Butterworth, S.; Crook, O. M.; Lassadi, I.; Gupta, V.; Tromer, E.; Mourier, T.; Stevens, T. J.; Breckels, L. M.; Pain, A.; Lilley, K. S.; Waller, R. F. A Comprehensive Subcellular Atlas of the Toxoplasma Proteome via hyperLOPIT Provides Spatial Context for Protein Functions. *Cell Host Microbe.* **2020**, *28*, 752−766.
(107) Mulvey, C. M.; Breckels, L. M.; Geladaki, A.; Britovsek, N. K.; Nightingale, D. J. H.; Christoforou, A.; Elzek, M.; Deery, M. J.; Gatto, L.; Lilley, K. S. Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat. Protoc.* **2017**, *12*, 1110−1135.
(108) Lund-Johansen, F.; de la Rosa Carrillo, D.; Mehta, A.; Sikorski, K.; Inngjerdingen, M.; Kalina, T.; Roysland, K.; de Souza, G. A.; Bradbury, A. R.; Lecrevisse, Q.; Stuchly, J. MetaMass, a tool for meta-analysis of subcellular proteomics data. *Nat. Methods.* **2016**, *13*, 837−840.
(109) Tan, D. J.; Dvinge, H.; Christoforou, A.; Bertone, P.; Martinez Arias, A.; Lilley, K. S. Mapping organelle proteins and protein complexes in Drosophila melanogaster. *J. Proteome Res.* **2009**, *8*, 2667−2678.
(110) Gatto, L.; Breckels, L. M.; Wieczorek, S.; Burger, T.; Lilley, K. S. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics.* **2014**, *30*, 1322−1324.
(111) Bannon, D.; Moen, E.; Schwartz, M.; Borba, E.; Kudo, T.; Greenwald, N.; Vijayakumar, V.; Chang, B.; Pao, E.; Osterman, E.; Graf, W.; Van Valen, D. DeepCell Kiosk: scaling deep learning-enabled cellular image analysis with Kubernetes. *Nat. Methods.* **2021**, *18*, 43−45.
(112) Weill, U.; Arakel, E. C.; Goldmann, O.; Golan, M.; Chuartzman, S.; Munro, S.; Schwappach, B.; Schuldiner, M. Toolbox: Creating a systematic database of secretory pathway proteins uncovers new cargo for COPI. *Traffic.* **2018**, *19*, 370−379.
(113) Snyder, M. P.; et al. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature.* **2019**, *574*, 187−192.
(114) Goltsev, Y.; Samusik, N.; Kennedy-Darling, J.; Bhate, S.; Hale, M.; Vazquez, G.; Black, S.; Nolan, G. P. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell.* **2018**, *174*, 968−981.
(115) Hasanaj, E.; Wang, J.; Sarathi, A.; Ding, J.; Bar-Joseph, Z. Interactive single-cell data analysis using Cellar. *Nat. Commun.* **2022**, *13*, 1998.
(116) Jakobsen, L.; Vanselow, K.; Skogs, M.; Toyoda, Y.; Lundberg, E.; Poser, I.; Falkenby, L. G.; Bennetzen, M.; Westendorf, J.; Nigg, E. A.; Uhlen, M.; Hyman, A. A.; Andersen, J. S. Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. *EMBO J.* **2011**, *30*, 1520−1535.
(117) Morgenstern, M.; Stiller, S. B.; Lubbert, P.; Peikert, C. D.; Dannenmaier, S.; Drepper, F.; Weill, U.; Hoss, P.; Feuerstein, R.; Gebert, M.; Bohnert, M.; van der Laan, M.; Schuldiner, M.; Schutze, C.; Oeljeklaus, S.; Pfanner, N.; Wiedemann, N.; Warscheid, B. Definition of a High-Confidence Mitochondrial Proteome at Quantitative Scale. *Cell Rep.* **2017**, *19*, 2836−2852.
(118) Mackinder, L. C. M.; Chen, C.; Leib, R. D.; Patena, W.; Blum, S. R.; Rodman, M.; Ramundo, S.; Adams, C. M.; Jonikas, M. C. A

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

Spatial Interactome Reveals the Protein Organization of the Algal CO2-Concentrating Mechanism. *Cell.* **2017**, *171*, 133−147.
(119) Jean Beltran, P. M.; Mathias, R. A.; Cristea, I. M. A Portrait of the Human Organelle Proteome In Space and Time during Cytomegalovirus Infection. *Cell Syst.* **2016**, *3*, 361−373.
(120) Lobingier, B. T.; Huttenhain, R.; Eichel, K.; Miller, K. B.; Ting, A. Y.; von Zastrow, M.; Krogan, N. J. An Approach to Spatiotemporally Resolve Protein Interaction Networks in Living Cells. *Cell.* **2017**, *169*, 350−360.
(121) Liu, X.; Salokas, K.; Tamene, F.; Jiu, Y.; Weldatsadik, R. G.; Ohman, T.; Varjosalo, M. An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nat. Commun.* **2018**, *9*, 1188.
(122) Chong, Y. T.; Koh, J. L.; Friesen, H.; Duffy, S. K.; Cox, M. J.; Moses, A.; Moffat, J.; Boone, C.; Andrews, B. J. Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell.* **2015**, *161*, 1413−1424.
(123) Breker, M.; Gymrek, M.; Schuldiner, M. A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *J. Cell Biol.* **2013**, *200*, 839−850.
(124) Tkach, J. M.; Yimit, A.; Lee, A. Y.; Riffle, M.; Costanzo, M.; Jaschob, D.; Hendry, J. A.; Ou, J.; Moffat, J.; Boone, C.; Davis, T. N.; Nislow, C.; Brown, G. W. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nat. Cell Biol.* **2012**, *14*, 966−976.
(125) Breckels, L. M.; Gatto, L.; Christoforou, A.; Groen, A. J.; Lilley, K. S.; Trotter, M. W. The effect of organelle discovery upon sub-cellular protein localisation. *J. Proteomics.* **2013**, *88*, 129−140.
(126) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.
(127) Fu, J.; Zhang, Y.; Wang, Y.; Zhang, H.; Liu, J.; Tang, J.; Yang, Q.; Sun, H.; Qiu, W.; Ma, Y.; Li, Z.; Zheng, M.; Zhu, F. Optimization of metabolomic data processing using NOREVA. *Nat. Protoc.* **2022**, *17*, 129−151.
(128) Yang, Q.; Wang, Y.; Zhang, Y.; Li, F.; Xia, W.; Zhou, Y.; Qiu, Y.; Li, H.; Zhu, F. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res.* **2020**, *48*, W436−W448.
(129) Chen, C.; Hou, J.; Tanner, J. J.; Cheng, J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int. J. Mol. Sci.* **2020**, *21*, 2873.
(130) Tang, J.; Fu, J.; Wang, Y.; Luo, Y.; Yang, Q.; Li, B.; Tu, G.; Hong, J.; Cui, X.; Chen, Y.; Yao, L.; Xue, W.; Zhu, F. Simultaneous Improvement in the Precision, Accuracy, and Robustness of Label-free Proteome Quantification by Optimizing Data Manipulation Chains. *Mol. Cell Proteomics.* **2019**, *18*, 1683−1699.
(131) Tang, J.; Fu, J.; Wang, Y.; Li, B.; Li, Y.; Yang, Q.; Cui, X.; Hong, J.; Li, X.; Chen, Y.; Xue, W.; Zhu, F. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform.* **2020**, *21*, 621−636.
(132) Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L. Improving label-free quantitative proteomics strategies by distributing shared peptides and stabilizing variance. *Anal. Chem.* **2015**, *87*, 4749−4756.
(133) Ringner, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303−304.
(134) Karimpour-Fard, A.; Epperson, L. E.; Hunter, L. E. A survey of computational tools for downstream analysis of proteomic and other omic datasets. *Hum Genomics.* **2015**, *9*, 28.
(135) Borner, G. H.; Hein, M. Y.; Hirst, J.; Edgar, J. R.; Mann, M.; Robinson, M. S. Fractionation profiling: a fast and versatile approach for mapping vesicle proteomes and protein-protein interactions. *Mol. Biol. Cell* **2014**, *25*, 3178−3194.
(136) Crook, O. M.; Geladaki, A.; Nightingale, D. J. H.; Vennard, O. L.; Lilley, K. S.; Gatto, L.; Kirk, P. D. W. A semi-supervised Bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. *PLoS Comput. Biol.* **2020**, *16*, No. e1008288.
(137) Trotter, M. W.; Sadowski, P. G.; Dunkley, T. P.; Groen, A. J.; Lilley, K. S. Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. *Proteomics.* **2010**, *10*, 4213−4219.
(138) Nikolovski, N.; Rubtsov, D.; Segura, M. P.; Miles, G. P.; Stevens, T. J.; Dunkley, T. P.; Munro, S.; Lilley, K. S.; Dupree, P. Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiol.* **2012**, *160*, 1037−1051.
(139) Dayao, M. T.; Brusko, M.; Wasserfall, C.; Bar-Joseph, Z. Membrane marker selection for segmenting single cell spatial proteomics data. *Nat. Commun.* **2022**, *13*, 1999.
(140) Van Valen, D. A.; Kudo, T.; Lane, K. M.; Macklin, D. N.; Quach, N. T.; DeFelice, M. M.; Maayan, I.; Tanouchi, Y.; Ashley, E. A.; Covert, M. W. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLoS Comput. Biol.* **2016**, *12*, No. e1005177.
(141) Berg, S.; Kutra, D.; Kroeger, T.; Straehle, C. N.; Kausler, B. X.; Haubold, C.; Schiegg, M.; Ales, J.; Beier, T.; Rudy, M.; Eren, K.; Cervantes, J. I.; Xu, B.; Beuttenmueller, F.; Wolny, A.; Zhang, C.; Koethe, U.; Hamprecht, F. A.; Kreshuk, A. ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods.* **2019**, *16*, 1226−1232.
(142) Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods.* **2012**, *9*, 671−675.
(143) Bray, M. A.; Fraser, A. N.; Hasaka, T. P.; Carpenter, A. E. Workflow and metrics for image quality control in large-scale high-content screens. *J. Biomol Screen.* **2012**, *17*, 266−274.
(144) Hulsman, M.; Hulshof, F.; Unadkat, H.; Papenburg, B. J.; Stamatialis, D. F.; Truckenmuller, R.; van Blitterswijk, C.; de Boer, J.; Reinders, M. J. Analysis of high-throughput screening reveals the effect of surface topographies on cellular morphology. *Acta Biomater.* **2015**, *15*, 29−38.
(145) Ramo, P.; Sacher, R.; Snijder, B.; Begemann, B.; Pelkmans, L. CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics.* **2009**, *25*, 3028−3030.
(146) Parnamaa, T.; Parts, L. Accurate Classification of Protein Subcellular Localization from High-Throughput Microscopy Images Using Deep Learning. *G3 (Bethesda).* **2017**, *7*, 1385−1392.
(147) Tang, J.; Mou, M.; Wang, Y.; Luo, Y.; Zhu, F. MetaFS: Performance assessment of biomarker discovery in metaproteomics. *Brief Bioinform.* **2021**, *22*, bbaa105.
(148) Li, F.; Zhou, Y.; Zhang, Y.; Yin, J.; Qiu, Y.; Gao, J.; Zhu, F. POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability. *Brief Bioinform.* **2022**, *23*, bbac040.
(149) Li, F.; Yin, J.; Lu, M.; Yang, Q.; Zeng, Z.; Zhang, B.; Li, Z.; Qiu, Y.; Dai, H.; Chen, Y.; Zhu, F. ConSIG: consistent discovery of molecular signature from OMIC data. *Brief Bioinform.* **2022**, *23*, bbac253.
(150) Ullah, M.; Han, K.; Hadi, F.; Xu, J.; Song, J.; Yu, D. J. PScL-HDeep: image-based prediction of protein subcellular location in human tissue using ensemble learning of handcrafted and deep learned features with two-layer feature selection. *Brief Bioinform.* **2021**, *22*, bbab278.
(151) Coelho, L. P.; Kangas, J. D.; Naik, A. W.; Osuna-Highley, E.; Glory-Afshar, E.; Fuhrman, M.; Simha, R.; Berget, P. B.; Jarvik, J. W.; Murphy, R. F. Determining the subcellular location of new proteins from microscope images using local features. *Bioinformatics.* **2013**, *29*, 2343−2349.
(152) Handfield, L. F.; Chong, Y. T.; Simmons, J.; Andrews, B. J.; Moses, A. M. Unsupervised clustering of subcellular protein expression patterns in high-throughput microscopy images reveals protein complexes and functional relationships between proteins. *PLoS Comput. Biol.* **2013**, *9*, No. e1003085.
(153) Newberg, J. Y.; Li, J.; Rao, A.; Ponten, F.; Uhlen, M.; Lundberg, E.; Murphy, R. F. Automated Analysis of Human Protein Atlas Immunofluorescence Images. *Proc. IEEE Int. Symp. Biomed Imaging.* **2009**, *5193229*, 1023−1026.

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

(154) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature.* **2015**, *521*, 436−444.
(155) Ouyang, W.; Winsnes, C. F.; Hjelmare, M.; Cesnik, A. J.; Akesson, L.; Xu, H.; Sullivan, D. P.; Dai, S.; Lan, J.; Jinmo, P.; Galib, S. M.; Henkel, C.; Hwang, K.; Poplavskiy, D.; Tunguz, B.; Wolfinger, R. D.; Gu, Y.; Li, C.; Xie, J.; Buslov, D.; Fironov, S.; Kiselev, A.; Panchenko, D.; Cao, X.; Wei, R.; Wu, Y.; Zhu, X.; Tseng, K. L.; Gao, Z.; Ju, C.; Yi, X.; Zheng, H.; Kappel, C.; Lundberg, E. Analysis of the Human Protein Atlas Image Classification competition. *Nat. Methods.* **2019**, *16*, 1254−1261.
(156) Greenwald, N. F.; Miller, G.; Moen, E.; Kong, A.; Kagel, A.; Dougherty, T.; Fullaway, C. C.; McIntosh, B. J.; Leow, K. X.; Schwartz, M. S.; Pavelchek, C.; Cui, S.; Camplisson, I.; Bar-Tal, O.; Singh, J.; Fong, M.; Chaudhry, G.; Abraham, Z.; Moseley, J.; Warshawsky, S.; Soon, E.; Greenbaum, S.; Risom, T.; Hollmann, T.; Bendall, S. C.; Keren, L.; Graf, W.; Angelo, M.; Van Valen, D. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **2022**, *40*, 555−565.
(157) McQuin, C.; Goodman, A.; Chernyshev, V.; Kamentsky, L.; Cimini, B. A.; Karhohs, K. W.; Doan, M.; Ding, L.; Rafelski, S. M.; Thirstrup, D.; Wiegraebe, W.; Singh, S.; Becker, T.; Caicedo, J. C.; Carpenter, A. E. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* **2018**, *16*, No. e2005970.
(158) Falk, T.; Mai, D.; Bensch, R.; Cicek, O.; Abdulkadir, A.; Marrakchi, Y.; Bohm, A.; Deubner, J.; Jackel, Z.; Seiwald, K.; Dovzhenko, A.; Tietz, O.; Dal Bosco, C.; Walsh, S.; Saltukoglu, D.; Tay, T. L.; Prinz, M.; Palme, K.; Simons, M.; Diester, I.; Brox, T.; Ronneberger, O. U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods.* **2019**, *16*, 67−70.
(159) Tu, Y.; Lei, H.; Shen, H. B.; Yang, Y. SIFLoc: a self-supervised pre-training method for enhancing the recognition of protein subcellular localization in immunofluorescence microscopic images. *Brief Bioinform.* **2022**, *23*, bbab605.
(160) Breckels, L. M.; Mulvey, C. M.; Lilley, K. S.; Gatto, L. A Bioconductor workflow for processing and analysing spatial proteomics data. *F1000Res.* **2016**, *5*, 2926.
(161) Zhang, C.; Mou, M.; Zhou, Y.; Zhang, W.; Lian, X.; Shi, S.; Lu, M.; Sun, H.; Li, F.; Wang, Y.; Zeng, Z.; Li, Z.; Zhang, B.; Qiu, Y.; Zhu, F.; Gao, J. Biological activities of drug inactive ingredients. *Brief Bioinform.* **2022**, *23*, bbac160.
(162) Carracedo-Reboredo, P.; Linares-Blanco, J.; Rodriguez-Fernandez, N.; Cedron, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. *Comput. Struct Biotechnol J.* **2021**, *19*, 4538−4558.
(163) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(164) Wu, Z.; Lei, T.; Shen, C.; Wang, Z.; Cao, D.; Hou, T. ADMET Evaluation in Drug Discovery. 19. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches. *J. Chem. Inf Model.* **2019**, *59*, 4587−4601.
(165) Yang, C.; Chen, J.; Wang, R.; Zhang, M.; Zhang, C.; Liu, J. Density Prediction Models for Energetic Compounds Merely Using Molecular Topology. *J. Chem. Inf Model.* **2021**, *61*, 2582−2593.
(166) Sullivan, D. P.; Winsnes, C. F.; Akesson, L.; Hjelmare, M.; Wiking, M.; Schutten, R.; Campbell, L.; Leifsson, H.; Rhodes, S.; Nordgren, A.; Smith, K.; Revaz, B.; Finnbogason, B.; Szantner, A.; Lundberg, E. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **2018**, *36*, 820−828.
(167) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf Model.* **2015**, *55*, 263−274.
(168) Eppenhof, K. A. J.; Lafarge, M. W.; Veta, M.; Pluim, J. P. W. Progressively Trained Convolutional Neural Networks for Deformable Image Registration. *IEEE Trans Med. Imaging.* **2020**, *39*, 1594−1604.
(169) Xia, W.; Zheng, L.; Fang, J.; Li, F.; Zhou, Y.; Zeng, Z.; Zhang, B.; Li, Z.; Li, H.; Zhu, F. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Comput. Biol. Med.* **2022**, *145*, 105465.
(170) Jing, L.; Tian, Y. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. *IEEE Trans Pattern Anal Mach Intell.* **2021**, *43*, 4037−4058.
(171) Kobayashi, H.; Cheveralls, K. C.; Leonetti, M. D.; Royer, L. A. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods.* **2022**, *19*, 995−1003.
(172) Stringer, C.; Wang, T.; Michaelos, M.; Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods.* **2021**, *18*, 100−106.
(173) Nadif, M.; Role, F. Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Brief Bioinform.* **2021**, *22*, 1592−1603.
(174) McKinley, E. T.; Shao, J.; Ellis, S. T.; Heiser, C. N.; Roland, J. T.; Macedonia, M. C.; Vega, P. N.; Shin, S.; Coffey, R. J.; Lau, K. S. MIRIAM: A machine and deep learning single-cell segmentation and quantification pipeline for multi-dimensional tissue images. *Cytometry A* **2022**, *101*, 521−528.
(175) Borner, G. H. H. Organellar Maps Through Proteomic Profiling - A Conceptual Guide. *Mol. Cell Proteomics.* **2020**, *19*, 1076−1087.
(176) Devarajan, K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.* **2008**, *4*, No. e1000029.
(177) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach Learn Res.* **2008**, *9*, 2579−2605.
(178) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C. A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **2019**, *37*, 38−44.
(179) Oyelade, J.; Isewon, I.; Oladipupo, F.; Aromolaran, O.; Uwoghiren, E.; Ameh, F.; Achas, M.; Adebiyi, E. Clustering Algorithms: Their Application to Gene Expression Data. *Bioinform Biol. Insights.* **2016**, *10*, 237−253.
(180) D'Haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **2005**, *23*, 1499−1501.
(181) Petegrosso, R.; Li, Z.; Kuang, R. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief Bioinform.* **2020**, *21*, 1209−1223.
(182) Rokach, L. Ensemble-based classifiers. *Artif Intell Rev.* **2010**, *33*, 1−39.
(183) Kraus, O. Z.; Ba, J. L.; Frey, B. J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics.* **2016**, *32*, i52−i59.
(184) Xiong, D.; Zhang, Z.; Wang, T.; Wang, X. A comparative study of multiple instance learning methods for cancer detection using T-cell receptor sequences. *Comput. Struct Biotechnol J.* **2021**, *19*, 3255−3268.
(185) Gut, G.; Herrmann, M. D.; Pelkmans, L. Multiplexed protein maps link subcellular organization to cellular states. *Science.* **2018**, *361*, No. eaar7042.
(186) Keren, L.; Bosse, M.; Marquez, D.; Angoshtari, R.; Jain, S.; Varma, S.; Yang, S. R.; Kurian, A.; Van Valen, D.; West, R.; Bendall, S. C.; Angelo, M. A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell.* **2018**, *174*, 1373−1387.
(187) Chandrasekaran, S. N.; Ceulemans, H.; Boyd, J. D.; Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discovery* **2021**, *20*, 145−159.
(188) Gatto, L.; Lilley, K. S. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics.* **2012**, *28*, 288−289.
(189) Kennedy, M. A.; Hofstadter, W. A.; Cristea, I. M. TRANSPIRE: A Computational Pipeline to Elucidate Intracellular

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895

Protein Movements from Spatial Proteomics Data Sets. *J. Am. Soc. Mass Spectrom.* **2020**, *31*, 1422−1439.
(190) Palla, G.; Spitzer, H.; Klein, M.; Fischer, D.; Schaar, A. C.; Kuemmerle, L. B.; Rybakov, S.; Ibarra, I. L.; Holmberg, O.; Virshup, I.; Lotfollahi, M.; Richter, S.; Theis, F. J. Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods.* **2022**, *19*, 171−178.
(191) Czech, E.; Aksoy, B. A.; Aksoy, P.; Hammerbacher, J. Cytokit: a single-cell analysis toolkit for high dimensional fluorescent microscopy imaging. *BMC Bioinformatics.* **2019**, *20*, 448.
(192) Dries, R.; Zhu, Q.; Dong, R.; Eng, C. L.; Li, H.; Liu, K.; Fu, Y.; Zhao, T.; Sarkar, A.; Bao, F.; George, R. E.; Pierson, N.; Cai, L.; Yuan, G. C. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **2021**, *22*, 78.
(193) Yan, W.; Hwang, D.; Aebersold, R. Quantitative proteomic analysis to profile dynamic changes in the spatial distribution of cellular proteins. *Methods Mol. Biol.* **2008**, *432*, 389−401.
(194) Crook, O. M.; Mulvey, C. M.; Kirk, P. D. W.; Lilley, K. S.; Gatto, L. A Bayesian mixture modelling approach for spatial proteomics. *PLoS Comput. Biol.* **2018**, *14*, No. e1006516.
(195) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y.; Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.
(196) Li, J.; Xiong, L.; Schneider, J.; Murphy, R. F. Protein subcellular location pattern classification in cellular images using latent discriminative models. *Bioinformatics.* **2012**, *28*, i32−i39.
(197) Jin, J.; Schorpp, K.; Samaga, D.; Unger, K.; Hadian, K.; Stockwell, B. R. Machine Learning Classifies Ferroptosis and Apoptosis Cell Death Modalities with TfR1 Immunostaining. *ACS Chem. Biol.* **2022**, *17*, 654−660.
(198) Jackson, H. W.; Fischer, J. R.; Zanotelli, V. R. T.; Ali, H. R.; Mechera, R.; Soysal, S. D.; Moch, H.; Muenst, S.; Varga, Z.; Weber, W. P.; Bodenmiller, B. The single-cell pathology landscape of breast cancer. *Nature.* **2020**, *578*, 615−620.
(199) Schulz, D.; Zanotelli, V. R. T.; Fischer, J. R.; Schapiro, D.; Engler, S.; Lun, X. K.; Jackson, H. W.; Bodenmiller, B. Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. *Cell Syst.* **2018**, *6*, 25−36.
(200) Dao, D.; Fraser, A. N.; Hung, J.; Ljosa, V.; Singh, S.; Carpenter, A. E. CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. *Bioinformatics.* **2016**, *32*, 3210−3212.
(201) Cao, Z. J.; Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **2022**, *40*, 1458.
(202) Kruse, A. R. S.; Spraggins, J. M. Uncovering Molecular Heterogeneity in the Kidney With Spatially Targeted Mass Spectrometry. *Front Physiol.* **2022**, *13*, 837773.
(203) Guilliams, M.; Bonnardel, J.; Haest, B.; Vanderborght, B.; Wagner, C.; Remmerie, A.; Bujko, A.; Martens, L.; Thone, T.; Browaeys, R.; De Ponti, F. F.; Vanneste, B.; Zwicker, C.; Svedberg, F. R.; Vanhalewyn, T.; Goncalves, A.; Lippens, S.; Devriendt, B.; Cox, E.; Ferrero, G.; Wittamer, V.; Willaert, A.; Kaptein, S. J. F.; Neyts, J.; Dallmeier, K.; Geldhof, P.; Casaert, S.; Deplancke, B.; Ten Dijke, P.; Hoorens, A.; Vanlander, A.; Berrevoet, F.; Van Nieuwenhove, Y.; Saeys, Y.; Saelens, W.; Van Vlierberghe, H.; Devisscher, L.; Scott, C. L. Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell.* **2022**, *185*, 379−396.
(204) Scupakova, K.; Soons, Z.; Ertaylan, G.; Pierzchalski, K. A.; Eijkel, G. B.; Ellis, S. R.; Greve, J. W.; Driessen, A.; Verheij, J.; De Kok, T. M.; Olde Damink, S. W. M.; Rensen, S. S.; Heeren, R. M. A. Spatial Systems Lipidomics Reveals Nonalcoholic Fatty Liver Disease Heterogeneity. *Anal. Chem.* **2018**, *90*, 5130−5138.

https://doi.org/10.1021/acs.jcim.2c01161
J. Chem. Inf. Model. 2022, 62, 5875−5895