SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(4), 432.
https://doi.org/10.62617/mcb432

Article

Application of spatial metrology models in cell molecular localization and functional prediction

Hongtao Wang*, Yulei Wang

Department of Basic Courses, Xinxiang Vocational and Technical College, Xinxiang 453006, Henan, China
*** Corresponding author:** Hongtao Wang, htwang361@126.com

CITATION

Wang H, Wang Y. Application of spatial metrology models in cell molecular localization and functional prediction. Molecular & Cellular Biomechanics. 2024; 21(4): 432. https://doi.org/10.62617/mcb432

ARTICLE INFO

Received: 27 September 2024
Accepted: 14 October 2024
Available online: 9 December 2024

Abstract: Understanding a protein's exact cellular location is often essential to understanding its function. Even with the advancements in computer approaches, protein localization prediction indeed faces major obstacles such as interpretability and handling numerous localization sites. In this research, a novel approach, Squirrel Search Optimized Dynamic Visual Geometry Group Network (SSO-DVGG), is proposed to improve protein sub-cellular localization predictions by utilizing spatial metrology models to tackle these problems. With its simplified architecture, SSO-DVGG can explain whether a protein is directed to particular cellular sites, as well as identify important sequence components like sorting motifs or localization signals. This model allows users to select acceptable error levels by providing a confidence estimate for each prediction and highlighting sequence properties that are responsible for localization. This makes the model interpretable. Furthermore, SSO-DVGG uses a probabilistic methodology and integrates a large amount of data from dual-targeted proteins, which enables it to predict multiple localization locations per protein accurately. SSO-DVGG outperforms the best predictors and shows superior capacity to predict multiple localizations when tested on several independent datasets. By providing a clear and accurate understanding of protein distribution and function, this method promotes the application of spatial metrology models in cell molecular localization and functional prediction.

Keywords: spatial metrology models; cell molecular localization; distribution and function; Squirrel Search Optimized Dynamic Visual Geometry Group Network (SSO-DVGG)

1. Introduction

The intricate landscape of cellular biology is defined by numerous molecular communications that strengthen necessary biological processes. Understanding the precise localization of these molecules within cells is critical for explaining their functional roles and the mechanisms underlying cellular behavior. Traditional methods for molecular localization often face limitations, including inadequate spatial resolution and challenges in quantifying dynamic connections within composite cellular environments [1]. A discipline traditionally rooted in engineering and physical sciences, offers a compelling framework for addressing these challenges. By utilizing advanced measurement techniques and modeling approaches, spatial metrology enables the accurate characterization of molecular positioning and spatial relationships at a scale relevant to cellular dynamics [2]. Biology has been completely transformed by optical microscopy of immune fluorescence-labeled materials, which makes it possible to observe cellular functions in their natural environment. However, due to the light diffraction along the optical path, an optical microscope's resolution is physically restricted to around 200 nm.

The organelles, cytoskeleton assemblies, and other scaffolding structures, such as Catherin-coated pits, which typically have dimensions between 10nm and 500 nm cannot be seen in detail because of the constraint. The restriction can currently be addressed by new optical methods known as super-resolution microscopy, which can resolve details as small as a few tens of nm. Super-resolution microscopy first appeared in the early 2000s, but since then, it has developed and is currently employed in many labs to study the nano-scale cellular architecture [3]. The intricate biological process of synaptic transmission, which facilitates neural communication, requires coordinating specialized protein complexes within interconnected cells. Neurotransmitters are released by proteins in the pre-synaptic active zone, which diffuses across the synaptic cleft and activates postsynaptic receptors. However, it is known about sub-synaptic structure and protein interactions despite their profound effects on synaptic transmission [4]. Particles and biological materials like cells and tissue can be characterized using EI. EI measurements have been used to ascertain the size and other fundamental particle characteristics and concentration, as well as to determine the type and viability of the cells. The ease of use, non-invasiveness, and lack of need for markers in simple sample characterization account for the popularity of EI sensing techniques. A further benefit of EI-based characterization techniques that adds to their wide range of applications is the abundance of inexpensive impedance measurement tools, such as lock-in amplifiers, and the ability to construct such apparatus in the lab from the ground up [5]. Mechanobiology studies how mechanical stimuli interact with cellular biology, encompassing how cells sense, transducer, and react to mechanical stimuli as well as how mechanical properties of cells are defined. Changes in mechanical pressures have the potential to cause remodeling of tissues, including blood vessels and bone, and have a substantial impact on cell behaviors and tissue homeostasis [6]. Furthermore, pattern creation, stem cell differentiation, adult tissue function, embryonic tissue development, and cell fate switching are all influenced by mechanical stresses. Current overviews published by various research teams offer more comprehensive information regarding mechanobiology and techniques for determining how cells react in response to mechanical forces such as expansion, and roughness of the surface [7]. The building blocks of all living things and biological processes are cells, which come in various forms, functions, and states. By examining the unique identity of every single cell, it is possible to analyze the interactions and composition of cells within biological systems and identify the subgroups that control biological functions. Genetic variety, which is phenol typically modified by genetic control or environmental disruption, is the source of cellular heterogeneity and manifests itself at several molecular levels [8]. Since the introduction of the concept, there has been debate on whether or not biological cells use quantum coherence for information processing. The debate gathered momentum when a remark argued that decoherence makes quantum computation in such systems impractical, in response to Hammer and Penrose's proposal that neural microtubules operate as quantum computers for further discussion. In quantum biology, recent research has mostly concentrated on the function of single-protein scale coherence in photoreception and magneto reception across a range of systems. There is yet a great deal of disagreement regarding theoretical interpretation and experimental replication. To assess

localization, two fluorescently labeled molecular species are typically compared and the overlap percentage between them is found [9, 10].

1.1. Research objective

The study aims to develop and validate a new model, SSO-DVGG, which is bound to be more accurate and interpretable for the protein predictions of cellular localization. Spatial metrology models are applied and a probabilistic methodology is integrated to overcome the current difficulties, such as the immense complexity involved in dealing with more than one localization site and thus identifying key components within the sequence that may correspond to sorting modifiers or localization signals. Thus, the study aims to increase the predictive performance and transparency of localization models, ultimately contributing to the knowledge of protein distribution and function.

1.2. Key contribution

- The SSO-DVGG model indeed outperforms all other protein localization predictions since it can predict more than one cellular location for a protein, increasing the reliability of overall predictions.
- The model provides confidence scores and identifies the crucial elements of sequence-like sorting determinants that help in understanding the reason why the prediction of localization is valid or not.
- By incorporating spatial metrology models, the study advances the methodology for protein localization prediction, enabling a more nuanced analysis of protein behavior within cellular environments.

1.3. Writing framework

The article is organized as follows: Section 2 offers relevant works, section 3 offers a thorough methodology, section 4 discusses the findings of the experiment, and Section 5 offers a conclusion.

2. Related work

To handle the impacts of variations with time in the mechanism of visual focusing, the CNN that has been used is trained over several days to take into consideration such variations. It detailed an ML-based optical autofocus system for microscopy that was long-range and reliable. It could have been useful in studies where long times were spent gathering image data, which could be affected by defocusing brought by component drift, like temperature changes or mechanical drift. Additionally, it aids in automatic slide scan and multiwall plates scan, where the samples to be scanned could move during the acquisition of image data from one horizontal plane to another, as described in [11]. Understanding the arrangement and clustering of proteins in a cell at the nanoscale was crucial for the onset of high-resolution image tools. The cluster analysis of SMLM images continues to be difficult since methods for SMLM image cluster analysis must be developed because standard computer cluster analysis techniques meant for regular with pointillism SMLM data and microscope images were not functional. It investigated how these

approaches have evolved by dividing computational cluster assessment techniques for SMLM images into two groups: ML-based and traditional [12]. Nearly single-cell resolution can be achieved in the identification of hundreds of metabolites in space using DESI and MALDI methods. The technology facilitated studies on the flexibility of cancer cells, tumor heterogeneity, and signals of communication between stromal cells and cancer in the tumor microenvironment. It was employed in clinical settings and has translational applications, such as determining drug distribution in organs and tumors [13]. The research focused on variations in transcription patterns among individual cells to illustrate how merely replicating cells, organs, and tissues must grow from single cells to multi-cellular ones. However, with the poor throughput and bulk levels of the research, it was difficult to comprehend the molecular mechanisms of differentiating in humans and animals. Fast progress in genomics has made it possible to perform genome-wide and finer-resolution investigations, especially with sc RNA-seq, which improved the comprehension of cell lineage and differentiation. Nevertheless, spatial information was lost with sc RNA-seq, which restricts comprehension of particular cell functions in various tissue locations [14]. The K2was a triple-color TIRF microscope that was available for free and was intended for imaging single molecules and live cells. Because of its modular architecture, users could enhance its functionality, adapt it to their needs, or reuse components to make improvements to current configurations. The K2 was intended to address the drawbacks of specially made microscopes, technical intricacy, and a dearth of publicly available instructions on how to construct customized configurations. Users could modify the K2 to suit their budgetary and scientific requirements and its modular design without having to construct a full duplicate [15]. It made sense to create a four-arm nanoprobe that could be used to image and identify different miRNAs in living cells more quickly and simultaneously. Acknowledgment of their complementary roles and guidance in the detection and treatment of human disorders, including cancer, has often been attributed to the development of very reliable and efficient tools for the simultaneous scanning of microRNAs in living cells [16]. Macromolecules in tomograms could be found using cryo-electron tomography (TM). However, computing constraints have restricted how far rotational searches could go in particle detection. The PyTOM software package included a GPU implementation of TM that enables sampling beyond the Crowther criterion and expedites orientation search. Automated extraction with high sensitivity was possible by sampling at the Crowther criteria. Low false-discovery rates facilitate automated ribosome classification using TM on locally adjusted tomograms using suitable angle sampling, allowing for high-resolution averaging and polysome structure identification [17]. To replicate realistic PSF forms in microscopy, especially for single-molecule localization techniques, a computer application has been developed. The tool could precisely simulate different imaging settings and included unique aberrations, transmission, and phase masks, as well as a variety of microscopy and fluorophore parameters. Moreover, it permitted the modeling of overlapped PSFs in dense molecule settings. To determine the best possible localization precision under specific circumstances, the program offered the Cramér-Rao bound. By generating a huge dataset with randomized simulation settings and enabling using experimental information to fit customized defects

straight, the software improved experimental design and result validation while closing the gap between experimental and simulated conditions [18]. Secreted signaling molecules were necessary for cells to carry out vital biological processes such as development, metabolism, and immunity. However, it was challenging to measure these activities with enough chemical specificity and temporal resolution. To allow for the continuous, real-time fluorescence monitoring of particular secreted analyses, researchers created an aptamer-conjugated hydrogel matrix. To study how camp signaling facilitated intercellular communication in Dictyosteliumdi sodium cells, an amoeba model, and real-time imaging were done. Aptamer switches were designed by the researchers to respond to camp signals by producing quick and reversible fluorescence shifts. The methodology could be expanded to quantify other released compounds to observe various extracellular signaling pathways and their physiological impacts in receiving cells [19]. A technique for utilizing nonlinear MTADS to estimate temperature and gas concentration in aircraft propulsion systems was detailed in [20]. Since inhomogeneous thermo physical variations were taken into account, the method produced non-zero mean fluctuations and non-Gaussian fluctuations in the spectrum absorbance output. Through the use of the inverse model, measuring resilience and reliability were increased in scarce photonic configurations.

Research gap

Significant advancements have been made in optical autofocus systems for microscopy, particularly with the integration of CNN to address thermal systems for long-term imaging of dynamic biological processes. The current methodologies often lack robustness in handling complex biological environments where fluctuations can lead to defocusing. Furthermore, the analysis of single SMLM data indeed faces challenges due to the inadequacy of traditional cluster analysis methods tailored for pointillism data. Despite the promising capabilities of super-resolution techniques, there was limited exploration of ML-based approaches for automated and accurate cluster analysis in SMLM. Additionally, while advancements in spatial genomics, such as sc RNA-seq, have provided detailed insights into cellular differentiation, they often sacrifice spatial context, complicating the understanding of cell functions within their native environments. Addressing these gaps can greatly increase the accuracy and reliability of microscopy techniques in the study of complex biological phenomena and facilitate more comprehensive insights into cellular behaviors and interactions, as shown in **Figure 1**.

The SSO-DVGG method addresses these issues by integrating spatial metrology models to enhance the interpretability and accuracy of protein localization predictions. Its simplified architecture allows for real-time adaptability to fluctuations, while probabilistic methodologies enable effective handling of multiple localization sites, ultimately improving reliability in dynamic biological environments. The **Figure 1** identifies several significant obstacles and unmet research needs in the field of sub-cellular localization forecasting, such as the lack of understanding regarding the sorting of protein molecules in organelles, the challenges associated with functional protein analysis, and the complexity of protein

relationships that call for sophisticated algorithms and extensive data analysis. Because of their complex organizational states, projecting multi-label and multiple-domain protein conformations is very time-consuming and costly, and there are still few reliable public tools for localization predictions. Furthermore, biochemical properties necessitate sophisticated, expensive, and inaccurate laboratory procedures, particularly when working with a large number of Gene Ontology terms. Informatics methods for localization predictions have advanced, however, problems with improving accuracy, handling multi-compartmental projections, and supplying operational descriptions for unidentified proteins still exist.

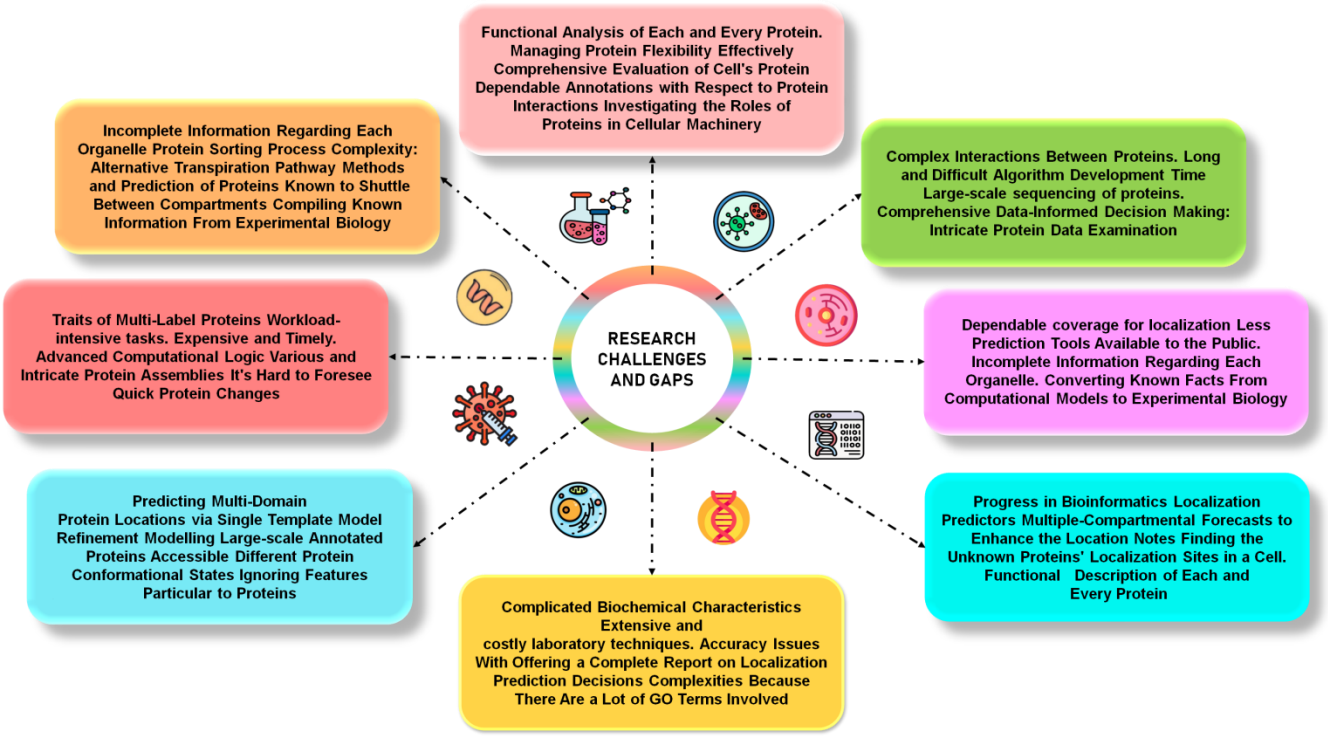

Figure 1. Sub-cellular localization prediction challenges and research gaps.

3. Methodology

The section addresses the framework and elements of SSO-DVGG to improve protein sub-cellular localization predictions by utilizing spatial metrology models to improve these problems. With its simplified architecture as shown in **Figure 2**, SSO-DVGG can explain whether a protein is directed to particular sub-cellular sites, as well as identify important sequence components like sorting motifs or localization signals. The primary steps of the process of machine learning are depicted in the diagram. The data collection phase, which involves gathering and preparing data, is where it starts. Pre-processing, which involves converting and cleaning the data to assure quality, is the following step. The next step involves feature extraction, which finds important characteristics to raise the precision of the model. Next is the prediction phase, in which the model creates forecasts or classifications using previously learned patterns. Lastly, evaluation of performance employs a variety of criteria to assess the model's efficacy and guarantee accuracy and dependability. Constructing a strong predictive model requires completing each step.

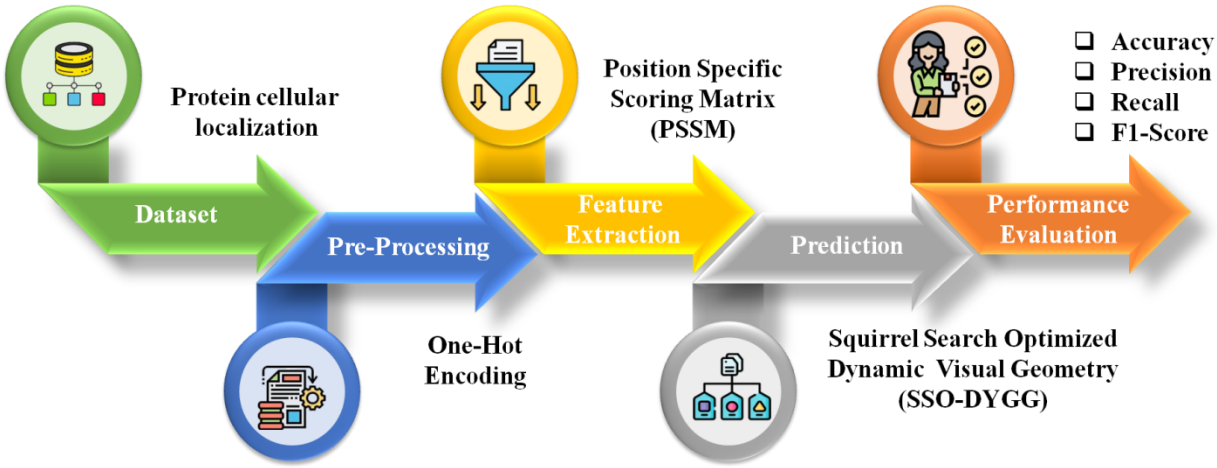

Figure 2. Block diagram of proposed flow.

3.1. Dataset

The regular structure of amino acids in molecules is analyzed to anticipate the likely physiological locations of peptides using protein sequence-based estimations of cellular translation. Pentapeptides, which are made up of five sequential amino acids, are important in this context since they are essential characteristics for classification. The protein sequencing is converted into mathematical models using a Count Vectorizer, which counts the instances of each pentapeptide in the sequences [21].

3.2. Pre-processing using one-hot encoding

Encoding techniques include one-hot encoding. The initial feature vector is expanded into a multidimensional matrix during one-hot encoding. The number of states in the feature determines the matrix's dimension, and each dimension corresponds to a distinct state. One benefit of one-hot encoding is its ability to remove the impact of digitized value differences on the model's training effect when digitizing classified-type features. More specifically, during the model-training process, a feature value encoded as 1000C can have a higher weight than one encoded as 1. One-hot encoding is used in the coding process to get rid of these kinds of negative effects. Moreover, one-hot encoding can address missing-value issues in the dataset by listing the missing values as a new dimension, which completes the missing dataset.

The fact that one-hot encoding offering does not attempt to imitate a missing value as a similar or calculated value gives it a significant edge over all other imputation techniques. Instead, it interprets the missing values as a separate class, which seeks another perspective to avoid the simulations interfering with data structure. The paper specifically processes the characteristics of such rules using one-hot encoding. Each distinct state value for a discrete feature is expanded to a new feature column to create a feature matrix. To create a feature matrix, the values of a continuous feature are first arranged in ascending order and then handled like a discrete feature. Moreover, each feature matrix's final column contains all missing values. Similar to imputation techniques, the training set is first rebuilt using one-hot

encoding, and the rule of dividing continuous features is then recorded and used in the testing set.

3.3. Feature extraction using Position-Specific Scoring Matrix (PSSM)

A popular feature expression type that offers extensive data on the evolution of protein sequences is PSSM. The structure and function of protein sequences are affected by the evolutionary information of proteins to a similar extent. Equation (1) can be used to represent the PSSM of each protein sequence and each amino acid in the sequence can be assigned a specified proportion in Equation (1).

$$O_{PSSM} = \begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,i} & \cdots & V_{1,20} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,i} & \cdots & V_{2,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{j,1} & V_{j,2} & \cdots & V_{j,i} & \cdots & V_{j,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{K,1} & V_{K,2} & \cdots & V_{K,i} & \cdots & V_{K,20} \end{bmatrix} \quad (1)$$

Each protein sequence's PSSM matrix is $L \times 20$, with L denoting the sequence's length, the 20 columns designating the different types of amino acids, and V denoting the amino acid position-specific score that causes a mutation from the protein sequence from the *j-th* position to the *i-th* place. The PSSM matrix elements are scaled by the sigmoid function to fall between 0 and 1, reducing bias and noise. The sigmoid function is expressed as follows in Equation (2).

$$e(w) = \frac{1}{1 + f^{-w}} \quad (2)$$

Here, $w$is the PSSM matrix element.

3.4. Prediction using Squirrel Search Optimized Dynamic Visual Geometry Group Network (SSO-DVGG)

The SSO-DVGG combines the strength of the SSA and the DVGG network to create a hybrid model that excels in protein sub-cellular localization prediction. SSA is utilized for optimizing the hyperparameters and searching for the best feature space, while DVGG is a deep learning technique. It is used because architecture can capture the spatial hierarchies found in protein sequence data. The combined strategy not only improves forecasting accuracy but also improves interpretability by highlighting key sequence components, such as sorting motifs and handling multiple localization sites simultaneously with high efficiency.

3.4.1. Dynamic Visual Geometry Group Network (DVGG)

Full-connected and multiple-connected convolutional layers make up the majority of each of the six main VGG CNN topologies. The input size is$224 \times 224 \times 3$, and the convolutional kernel has dimensions of3×3. Most of the time, they are between 16 and 19 layers. **Figure 3** displays the structure of the VGG-19 framework. VGG-19 CNN serves as a paradigm for pre-processing. Compared to standard CNN, the network depth has improved. The term DVGG describes a sophisticated architecture for neural networks that expands upon the Visual Geometry Group (VGG) models, which are widely utilized for the recognition of

objects and image classification applications because of extremely deep layers of convolution. By adding dynamic processes like temporal characteristics or adaptive learning, DVGG improves on the original VGG to handle and evaluate visual data more efficiently, especially in situations where modifications to visual information over time are significant. With its dynamic capacity, DVGG offers a more potent and adaptable method of processing visual information, making it an excellent choice for jobs like analyzing videos, real-time monitoring of objects, or other situations where visual trends change over time.

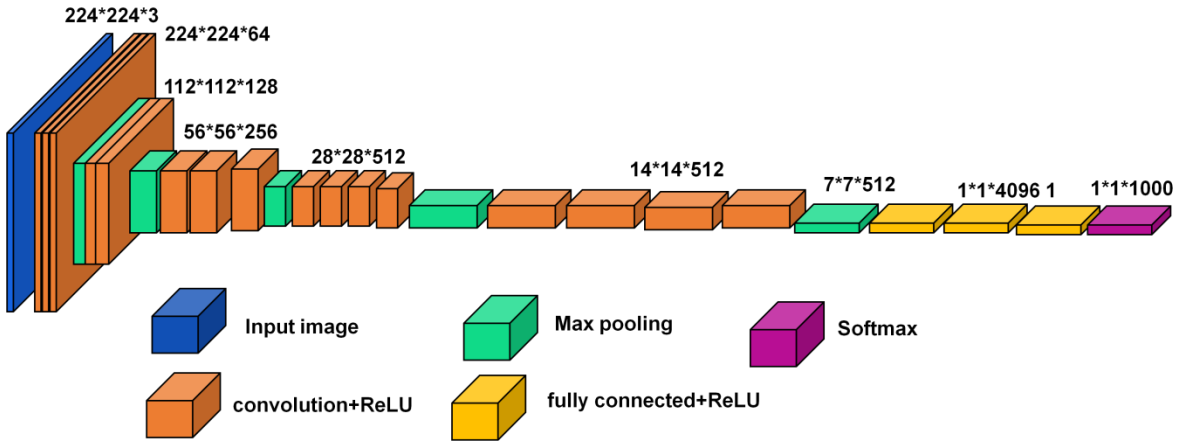

Figure 3. Flow diagram of DVGG.

Due to its alternating structure of many convolutional layering and irregular activation layers, it is better than utilizing a single convolution. The layer structure can shift the stimulus factor to the linear unit (ReLU), extract image features more successfully, and choose the largest value in an image region to be the pooled value for that area. Down sampling is accomplished by maxpooling. The down-sampling layer's primary objective is to minimize the number of parameters while preserving the sample's key characteristics, all while strengthening the network's ability to withstand visual distortion. Among these, τ_i^misthe coefficient for the i^{th}, the ReLU activation function is represented by$m$ in the feature map of the M^{th} layer given by Equation (3).

$$(\tau_i^m down\,(W_j^{m-1}) + a_j^{(m)}) \tag{3}$$

The maximum pooling sampling function is $down\,(W_j^{m-1})$.

3.4.2. Squirrel Search Optimization (SSO)

The process of hunting commences when flying squirrels start to scavenge. Fall brings the squirrels skimming from tree to tree in search of food sources. They move around and explore different areas of the woods at the same time. They eat oak seeds readily available since the climate is hot enough for them to complete their daily energy needs more quickly. As a result, they immediately consume oak seeds after learning about them. After meeting their daily energy needs, they start looking for the best wintertime food source, which happens to be hickory nuts. Hickory nuts' capacity will enable them to maintain their energy requirements in extremely harsh weather, cut down on costly search and rescue trips, and increase the likelihood of

survival. In deciduous woodlands, the loss of leaves throughout the winter months increases the risk of predation, making the animals less active but not hibernating. As winter draws to an end, flying squirrels regain their dynamic nature. A one-time process that structures the establishment of SSO and continues until the flying squirrel reaches its life expectancy. The SSO updates squirrel locations based on the type of squirrels, the SSO and flow of the season, and the appearance of chasers, as shown in **Figure 4**.

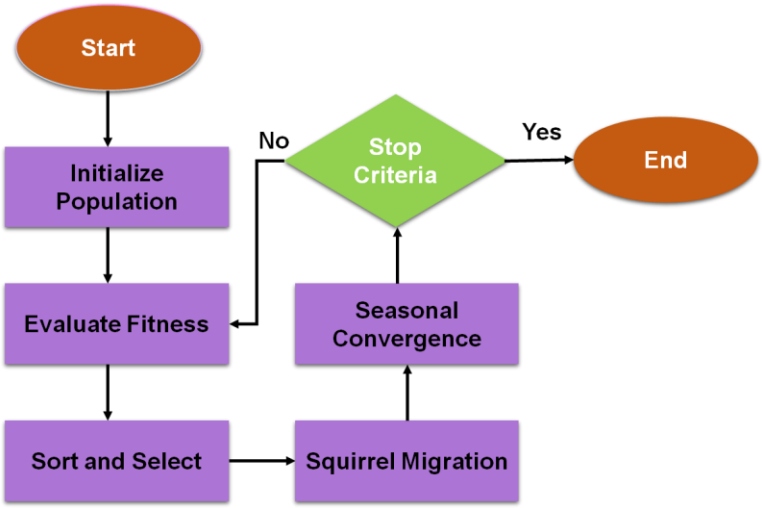

Figure 4. Flow chart of SSO.

Establish the populace

The upper and lower bounds of the pursuit space areW_V *and* W_k, assuming that the population is$M$. The following is an arbitrary creation of Msquirrels in Equation (4).

$$W_j = W_K + rand\ (1, D) \times (W_V - W_k) \tag{4}$$

where D is the issue's measurement, W_j denotes the i^{th}squirrel$(j = 1{:}M)$, and rand is a random value between 0 and 1.

Sort the populace

Assuming that there is only one squirrel per tree and that there are N total squirrels in the woods, SSO demands that there be one squirrel per tree. There is one hickory tree and one oak seed tree among all the N trees, the other trees are ordinary trees that receive no sustenance. The hickory tree provides the squirrels with the best nutrition, with the oak seed tree coming in second. Sorting the squirrels into three categories based on population fitness estimates in ascending order.

- A group of squirrels near hickory trees (x_g)
- Squirrels in the vicinity of oak seed trees (X_a)
- Squirrels positioned in typical trees(W_n)

Update the squirrels' location

The squirrels reassess their circumstances by skimming to the oak seed trees or hickory trees, as shown below.

$$W_j^{s+1} = \begin{cases} w_j^s + c_h H_d (W_{bj}^s - W_j^s & if\ q_1 \geq O_{co} \\ Rando\ location & otherwise \end{cases} \quad (5)$$

$$W_j^{s+1} = \begin{cases} w_j^s + c_h H_d (W_{bj}^s - W_j^s & if\ q_2 \geq O_{co} \\ Rando\ location & otherwise \end{cases} \quad (6)$$

The chaser likelihood is indicated byO_{co}, which is valued at 0.1. When $q_1 \geq O_{co}$, no chaser appears in Equation (5), allowing the squirrels to coast through the forest to find food and remain safe when $q_2 \geq O_{co}$, chasers appear, forcing the squirrels to restrict the amount of exercise they do, putting them in danger and causing them to move arbitrarily in Equation (6).$C_h$is possible to determine the skimming separation by Equation (7).

$$C_h = \frac{g_h}{\tan \emptyset} \quad (7)$$

In this instanceg_h is the consistent esteemed 8; $\tan \emptyset$indicates the coasting point that can be determined by Equation (8).

$$\tan \emptyset = \frac{C}{K} \quad (8)$$

The following can be used to estimate the lift and drag powers in Equations (9) and (10).

$$C = \frac{1}{2\rho u^2 T D_C} \quad (9)$$

$$K = \frac{1}{2\rho u^2 T D_K} \quad (10)$$

Random refreshes and sporadic verdict changes

The SSO requires that all of the population be in winter at the beginning of each generation, meaning that all of the squirrels must be fed by Equations (11) and (12), regardless of whether the season change is determined by the following equations.At the moment, the squirrels are refreshed.

$$T_d^s = \sqrt{\sum_{l=1}^{c} (W_{bj,L}^s - W_{g,l}^s)^2} \quad j = 1,2,\ldots,M_b \quad (11)$$

$$T_{min} = \frac{10 f^{-6}}{(365)^2 (\frac{S max}{2.5})} \quad (12)$$

The season remains the same if$T_{sd} < T_{min}$, meaning that winter has ended and summer has begun. When the season changes to summer, every individual that floats to W_h stays in the refreshed area, and every squirrel that skims to W_b without encountering any chasers adjusts their circumstances in the manner described below in Equations (13) and (14).

$$W_{jnew}^{s+!} = w_k + le'vy(w) \times (W_V - W_k) \quad (13)$$

$$Le'vy(w) = 0.01 \times \frac{\propto \times Q_b}{|Q_c|^{\frac{1}{\beta}}} \tag{14}$$

Levy is the arbitrary walk model, whose progression can be ascertained by and conforms to the Le'vy appropriation. ∝is determined as Equation (15).

$$\propto = \left[\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma(1+\beta) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}}\right] \tag{15}$$

The hybrid prediction SSO-DVGG integrates the SSA with the VGG-19 model. SSA optimizes feature selection by mimicking squirrel foraging behavior, while VGG-19 deep architecture features extraction with its layer's convolutional structure. The combination improves prediction accuracy, particularly for complex tasks like protein cellular localization, by maximizing computational overhead with effective hyperparameter tuning and efficient down-sampling techniques.

3.5. Spatial metrology models

Spatial metrology models are advanced techniques used to measure and analyze spatial data, especially in fields requiring high precision, such as engineering manufacturing, and biological sciences. These models are fundamental to capturing such intricate spatial arrangements between objects and structures, often at a micro or Nanoscale. It applies a combination of mathematical algorithms and imaging technologies, like 3D scanning, to produce the most accurate geometrical models of an object that can be applied to map cellular structures, which allows for such precise localization and even predictions of function within bimolecular structures inside cells. Importantly, such models support predictive analysis, error minimization, and quality control in manufacturing through accurate measurements of critical components, alignment, and detection of deviations. AI-based spatial metrology models enhance real-time processing and improve interpretability, and hence are extremely potent in a complete range of applications: Semiconductor fabrication, cellular molecular localization in protein research, and more.

Therefore, the SSO-DVGG is merged with the advantages provided by spatial metrology models to enhance protein localization predictions, as it will make applications based on exact spatial data analysis possible. In this sense, SSO-DVGG combines advanced algorithms with technologies in order to accurately capture the detailed spatial relationship at micro scales, and this highly improves the accuracy of bimolecular mapping. The hybrid approach also makes it invaluable, especially when it comes to applications because it not only offers real-time predictive analysis but also considerably enhances interpretability.

The SS-DVGG is shown in Algorithm 1 which outlines the main step of the SSO-DVGG method and integer's spatial metrology aspects in predicting protein localization.

Algorithm 1 SS-DVGG

1. Initialize parameters:
2. Set population size (M)
3. Define bounds W_k and W_V for the search space
4. Initialize Squirrel population $W[j]\, for j\, =\, 1\, to M$
5. For each squirrel $j$in W:
6. Randomly assign position:
7. $W[j] = W_l + rand(1, D) \times (W_u - W_l)$
8. Evaluate the fitness of each squirrel based on protein localization accuracy.
9. Sort squirrels into three categories:
10. x_g: Best squirrels near hickory trees
11. X_a: Squirrels near oak seed trees
12. W_n: Ordinary squirrels
13. Update squirrel positions:
14. For each squirrel $jinW$:
15. If random value q1 ≥ O_co (0.1):
    $W[j] = W[j] + C_h \times H_d \times (W_{bj} - W[j])$
16. =Else:
17. $W[j] = Random\, location$
18. Apply down sampling and feature extraction using VGG-19 architecture:
19. Input image size: $224x224x3$
20. Convolution with kernel size $3x3$
21. Use the ReLU activation function and max pooling for downsampling.
22. Predict protein localization:
23. Utilize spatial metrology models to refine predictions based on spatial relationships.
24. Output results:
25. Display protein localization predictions with confidence estimates.
26. Repeat steps 3 to 8 for defined iterations or until convergence criteria are met.
27. End

Advantages of spatial metrology models

- Measurement in Space To locate molecular elements within cellular architecture with great precision and to precisely map cell processes, models make use of geometrical and spatial data.
- Recognizing complicated biological relationships and movements requires the ability to integrate data from different scales efficiently reflecting local as well as worldwide patterns inside cellular settings. This is made possible by such models.
- Geographical measuring models provide a greater understanding of biological events by enabling complete studies that improve forecasts of molecular relationships and cellular activities by fusing spatial specificity with operational data.

The suggested Cell Molecular Identification utilizing Spatial Metrology Models is superior to the existing techniques, such as Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) are useful tools for examining temporal patterns and complicated data payments, respectively, they might not adequately represent the complex spatial interactions present in cellular settings. More sophisticated comprehension of spatial trends and relationships among molecular elements is made possible by the application of spatial metrology for cell molecular identification. This method improves functional projections and integrates rich geometry data to improve precise localization. The spatial metrology models in SSO-DVGG techniques outperform other models in terms of efficiently recognizing and interpreting cellular elements and their actions.

4. Result

4.1. Experimental setup

In the study, it employed 3.11.4 to implement and execute the required algorithms. The experiments were conducted on a high-performance desktop configured with 64 GB RAM and an AMD Ryzen 5900X processor running Windows 11. The setup provided enhanced computational capabilities, facilitating the evaluation of the proposed optimization techniques under various conditions and ensuring reliable performance during extensive simulations.

4.2. Parameters setup

The hypermeters for SSO-DVGG and spatial metrology methods are described in **Table 1**.

Table 1. Parameters setup.

Hyperparameters	**Typical values**
Hidden units per dense layer	128, 256, 512
Epochs	50,100
Dropout rate	0.3,0.5,0.6
Optimizer	SSO, DVGG, spatial metrology
Batch size	32,64,128
RMSprop$\beta - 2$	0.999,0.99
Momentum$(\beta - 1)$	0.9, 0.95
Number of filters	32,64, 128
Number of Convolutional layers	2,3
Filter size	3 × 3, 5 × 5
Number of dense layers	2, 3
Activation function	ReLU, Leaky
Gradient accumulation steps	1, 4
Learning rate	0.001,0.0001,0.01
Pooling Size	2 × 2, 3 × 3
Spatial resolution (metrology)	1μm,500 nm
Measurement precise	0.1 nm, 1nm
Sampling rate	100Hz, 1 KHx
Error margin	1%, 5%

The **Table 1** lists the hyperparameters that are used and associated common values, especially when it comes to neural networks and spatial metrology, that are utilized to optimize predictive models for proteins cellular distribution predictions. The number of training epochs (usually set to 50 or 100), the quantity of undetectable units per dense layer (varying from 128 to 512), and different dropout rates (0.3 to 0.6) to avoid overestimation are important parameter sets. It also includes a list of several optimization techniques, including spatial measurement, SSO-DVGG, as well as the quantities of batches (32, 64, and 128) and

comprehension rates (0.001, 0.0001, and 0.01) that regulate the training procedure. The number of filters and convolutional layers, filter widths, activation functions (such as ReLU and Leaky ReLU), differential aggregation and combining size environments, and other important factors all affect how well the algorithm can learn intricate patterns. Additionally, for physical measurement programs, resolution of space and measuring accuracy are crucial; figures like 1μm or 500nm, along with tolerances for error of 1% or 5%, demonstrate how well the model localizes molecules at cellular levels. This thorough rundown of hyperparameters is crucial for optimizing models to improve their ability to forecast in biological environments.

4.3. Evaluation phase

The effectiveness of a suggested strategy is evaluated based on protein localization as the Endoplasmic Reticulum (ER), Golgi, Lysosome, Nucleus, Cytoplasm, Plasma Membrane, and Mitochondria examining performance metrics including f1-score, recall, accuracy, and precision.

4.3.1. Accuracy

The percentage of accurate predictions, which makes the model relative to all forecasts, is known as accuracy. When it comes to protein localization, the model's accuracy in determining a protein's cellular location is measured across all examined samples. It means how well the model locates and anticipates the exact position and purpose of chemicals within neurons. It gauges how well the actual biological information matches the predicted structural roles and locations. Dependable spatial and operational insights in biological settings are ensured by high accuracy, which is essential for the advancement of biological investigation and application.

Table 2 and **Figure 5** present the accuracy of the model in predicting protein localization across various cellular sites. The model shows high accuracy in identifying proteins localized to the Nucleus (0.90), Golgi (0.88), and Cytoplasm (0.87), indicating strong performance in predicting these sites. The ER and mitochondria also exhibit fairly high accuracy at 0.85 and 0.86, reflecting reliable perditions for these organelles. Lower but rather substantial accuracy values are observed for the Plasma Membrane (0.83) and Lysosome (0.82), suggesting slightly more challenges in predicting proteins localized to these areas. Overall, the model demonstrates robust performance across all localization sites, with accuracy values consistently above 0.80.

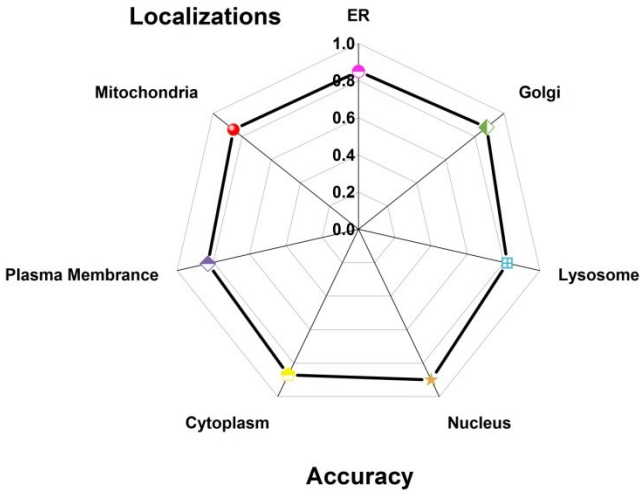

Figure 5. Outcome of accuracy.

Table 2. Results of accuracy.

Localizations	Accuracy
ER	0.85
Golgi	0.88
Lysosome	0.82
Nucleus	0.90
Cytoplasm	0.87
Plasma Membrane	0.83
Mitochondria	0.86

4.3.2. Precision

Precision can be defined as the actual optimistic forecast ratio divided by all predicted optimistic predictions. It demonstrated how well the model predicted a particular localization location. High precision in the study indicates that the model is probably accurate when it predicts a protein will be localized to a particular location. The reliability that occurs when the simulation accurately determines the actual chemical situations within a cell is referred to as precision. To reduce false positives, it calculates the percentage of correctly anticipated localizations (also known as true positives) among all predicted localities. The model's capacity to accurately identify certain molecular roles and locations within cellular settings is ensured by its high specificity.

The precision values in **Table 3** and **Figure 6** represent the values for different molecule localizations are summarized in the table, which shows how accurate the model is in recognizing specific proteins within particular parts of cells. With a precision of 0.88, the Nucleus exhibits the highest level of achievement, closely followed by the Golgi at 0.85. Additional localizations that exhibit strong predictive power include the Plasma Membrane (0.8) and the Cytoplasm (0.84). The Lysosome, on the other hand, has a somewhat lower precision of 0.78, indicating possible areas where the model's predictions should be strengthened.

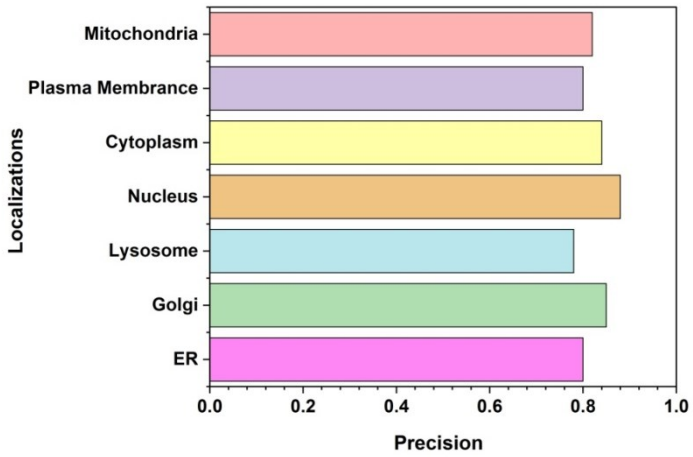

Figure 6. Outcome of precision.

Table 3. Results of precision.

Localizations	Precision
ER	0.8
Golgi	0.85
Lysosome	0.78
Nucleus	0.88
Cytoplasm	0.84
Plasma Membrane	0.8
Mitochondria	0.82

4.3.3. Recall

Recall, which is often referred to as sensitivity, quantifies the percentage of actual positives as opposed to genuine positive forecasts. It illustrates how well the model can detect proteins that are localized to a certain location. High recall in the study showed that the model correctly detects the majority of proteins that ought to be present at the spot. Recall gauges how well a model can pinpoint pertinent molecular positions or functionality predictions inside a cell. It shows the percentage of real molecular locations or functionalities that the simulation correctly identified. A high recall rate means that the majority of the real spatial or functional structures inside the cell are correctly captured by the model.

The recall values in **Table 4** and **Figure 7** represent the model's accuracy in predicting specific protein localizations. A higher recall like 0.85 Nucleus, indicated fewer false positives, meaning the model is more reliable in predicting that a protein is localized there. Lower values, like 0.78 for the ER and Plasma Membrane, suggested slightly less confidence in predictions for that site.

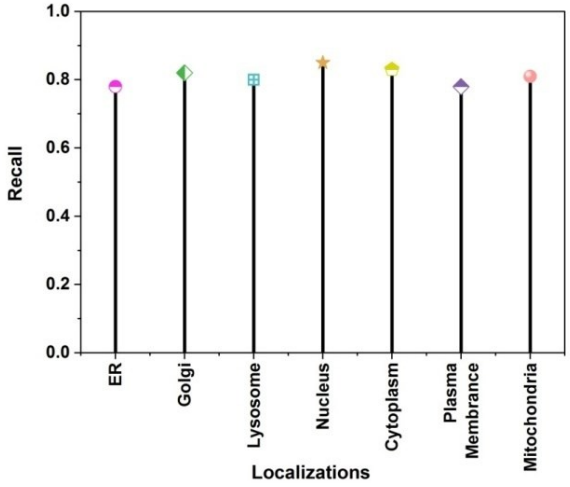

Figure 7. Outcome of recall.

Table 4. Results of recall.

Localizations	Recall
ER	0.78
Golgi	0.82
Lysosome	0.80
Nucleus	0.85
Cytoplasm	0.83
Plasma Membrane	0.78
Mitochondria	0.81

4.3.4. F1-score

Calculated as the harmonic mean of the two, the F1-score is a single metric that balances recall and precision. When predicting numerous localization sites or other scenarios with an unequal class distribution, it is very helpful. The F1-score measures the model's overall efficacies in protein localization prediction, taking into account both false positives and false negatives. Because it offers a single metric that accounts for both incorrect results and incorrect findings in the prediction of cellular activities or molecular locations, it is especially helpful in cases where the dataset is unbalanced. The efficacy of the model in accurately and thoroughly mapping components in cell localization duties is indicated by a high F1 score.

The values of the F1-score in **Table 5** and **Figure 8** represent, for each protein localization, the accuracy of model predictions. A high recall score, for instance, with 0.86 for the Nucleus, shows a low number of false positives, while the model is more reliable in predicting it to be located at that site. Low values, such as 0.78 for the ER, point out that, comparatively, there is less confidence in predictions about that site.

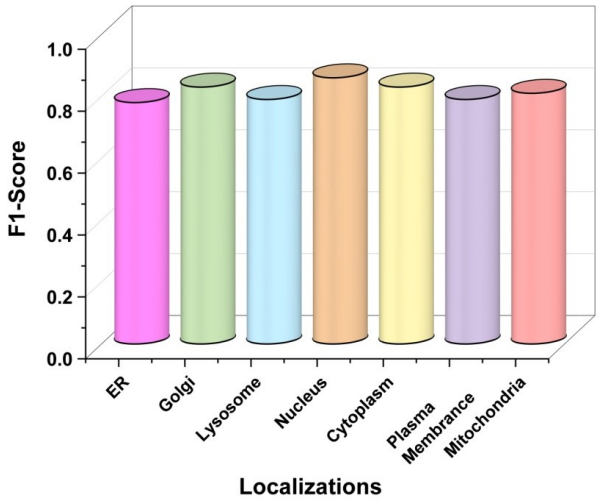

Figure 8. Outcomes of F1-score.

Table 5. Results of F1-score.

Localizations	F1-Score
ER	0.78
Golgi	0.83
Lysosome	0.79
Nucleus	0.86
Cytoplasm	0.83
Plasma Membrane	0.79
Mitochondria	0.81

5. Conclusion

The study aimed to advance the understanding of protein cellular localization by developing the SSO-DVGG, which utilizes spatial metrology models to enhance prediction accuracy and interpretability. The objective was to address challenges related to handling multiple localization sites and providing confidence estimates for prediction. By employing a simplified architecture, the SSO-DVGG model effectively predicted protein localization while highlighting essential sequence components such as sorting motifs. The findings demonstrate that the SSO-DVGG model achieves an accuracy of 0.90 across a variety of sections, including the ER, Golgi, Lysosome, Nucleus, Cytoplasm, Plasma Membrane, and Mitochondria, thereby considerably improving the prediction of diverse protein localizations. Among all metrics, the nucleus notably shows the highest values: 0.88 for precision, 0.85 for recall, and 0.86 for F1-score. This illustrates how well the model represents and localizes components in the nucleus in contrast to other cells and their compartments.

Limitations and future scope

This is a limitation, however, as the publicly available datasets used for training the model will not represent any possible localization situation. So, its applicability

might be somewhat limited to less characterized proteins. Future studies can explore incorporating all types of biological data and increase the size of the training dataset, thus making it a more diverse one to thereby enhance its generalization capabilities. In addition, the work that is being carried out on the SSO-DVGG in the bioinformatics area of function prediction and on investigations into interactions could hold very important information on protein dynamics and cellular processes. Appendix 1 depicts the List of Abbreviations.

Author contributions: Conceptualization, HW and YW; writing—original draft preparation, HW and YW; writing—review and editing, HW and YW. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Morris AR, Stanton DL, Roman D, Liu AC.Systems-level understanding of circadian integration with cell physiology. Journal of molecular biology. 2020; 432(12): 3547–3564.
2. McFarlane A, Pohler E, Moraga I. Molecular and cellular factors determine the functional pleiotropy of cytokines. The FEBS Journal. 2023; 290(10): 2525–2552.
3. Jimenez A, Friedl K, Leterrier C. About samples, giving examples: optimized single molecule localization microscopy. Methods. 2020; 174: 100–114.
4. Chen JH, Blanpied TA, Tang AH. Quantification of trans-synaptic protein alignment: A data analysis case for single-molecule localization microscopy. Methods. 2020; 174:72–80.
5. Schwarz M, Jendrusch M, Constantinou I. Spatially resolved electrical impedance methods for cell and particle characterization. Electrophoresis. 2020; 41(1–2): 65–80.
6. Hao Y, Cheng S, Tanaka, et al. Mechanical properties of single cells: Measurement methods and applications. Biotechnology Advances. 2020; 45: 107648.
7. Paul I, White C, Turcinovic I, Emili A. Imaging the future: the emerging era of single-cell spatial proteomics. The FEBS journal. 2021; 288(24): 6990–7001.
8. Chen Y, Song J, Ruan Q, et al.Single-cell sequencing methodologies: from transcriptome to multi-dimensional measurement. Small Methods. 2021; 5(6): 2100111.
9. Cheng MHY, Mo Y, Zheng G. Nano versus molecular: Optical imaging approaches to detect and monitor tumor hypoxia. Advanced Healthcare Materials. 2021; 10(2): 2001549.
10. Pelicci S, Furia L, Scanarini M, et al. Novel tools to measure single molecules colocalization in fluorescence nanoscopy by image cross-correlation spectroscopy. Nanomaterials. 2022; 12(4): 686.
11. Lightley J, Görlitz F, Kumar S, et al. Robust deep learning optical autofocus system applied to automated 20ultiwall plate single molecule localization microscopy.Journal of Microscopy. 2022; 288(2): 130–141.
12. Hyun Y, Kim D. The recent development of computational cluster analysis methods for single-molecule localization microscopy images. Computational and Structural Biotechnology Journal. 2023; 21: 879–888.
13. Planque M, Igelmann S, Campos AMF, Fendt SM. Spatial metabolomics principles and application to cancer research. Current Opinion in Chemical Biology. 2023; 76: 102362.
14. Choe K, Pak U, Pang Y, et al. Advances and challenges in spatial transcriptomics for developmental biology. Biomolecules. 2023; 13(1), p.156.
15. Niederauer C, Seynen M, Zomerdijk J, et al. The K2: Open-source simultaneous triple-color TIRF microscope for live-cell and single-molecule imaging. HardwareX. 2023; 13, p.e00404.
16. Xu H, Zheng Y, Fang X, et al. Spatial confinement-based figure-of-eight nanoknots accelerated simultaneous detection and imaging of intracellular microRNAs. AnalyticaChimicaActa. 2023; 1250, p.340974.

17. Chaillet ML, van der Schot G, Gubins I, et al. Extensive angular sampling enables the sensitive localization of macromolecules in electron tomograms. International Journal of Molecular Sciences. 2023; 24(17), p.13375.
18. Schneider MC, Hinterer F, Jesacher A, Schütz GJ. Interactive simulation and visualization of the point spread functions in single-molecule imaging. Optics Communications. 2024; 560, p.130463.
19. Park CH, Thompson IA, Newman SS, et al. Real-time spatiotemporal Measurement of Extracellular Signaling Molecules Using an Aptamer Switch-Conjugated Hydrogel Matrix. Advanced Materials. 2024; 36(4), p.2306704.
20. Niu Z, Qi H, Zhu Z, et al. Nonlinear multispectral tomographic absorption deflection spectroscopy based on Bayesian estimation for spatially resolved multiparameter measurement in methane flame exhaust. Fuel.2024; 357, p.129981.
21. Available online:https://www.kaggle.com/datasets/swlew369/protein-locations(accessed on 15 June 2023).

Appendix

EI	**electrical impedance**
CNN	Convolutional neural network
ML	machine learning
SMLM	single-molecule localization microscopy
DESI	Desorption Electrospray Ionization
MALDI	Matrix Assisted Laser Desorption Ionization
siRNA-seq	single-cell RNA-sequencing
TIRF	total internal reflection fluorescence
SPACIAL-CHA	spatial confinement-based dual-catalytic hairpin assembly
TM	template matching
PSF	point spread function
MTADS	multispectral tomographic absorption deflection spectroscopy
MAP	maximum a posteriori
NM	nanometers
K2	Kappa2
miRNA	Micro RNA
GPU	Graphics processing unit
cAMP	Cyclic adenosine monophosphate
AI	Artificial intelligence
3D	Three-dimensional