CAPSUL: A COMPREHENSIVE HUMAN PROTEIN BENCHMARK FOR SUBCELLULAR LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Subcellular localization is a crucial biological task for drug target identification and function annotation. Although it has been biologically realized that subcellular localization is closely associated with protein structure, no existing dataset offers comprehensive 3D structural information with detailed subcellular localization annotations, thus severely hindering the application of promising structure-based models on this task. To address this gap, we introduce a new benchmark called **CAPSUL**, a **C**omprehensive hum**A**n **P**rotein benchmark for **SU**bcellular **L**ocalization. It features a dataset that integrates diverse 3D structural representations with fine-grained subcellular localization annotations carefully curated by domain experts. We evaluate this benchmark using a variety of state-of-the-art sequence-based and structure-based models, showcasing the importance of involving structural features in this task. Furthermore, we explore reweighting and single-label classification strategies to facilitate future investigation on structure-based methods for this task. Lastly, we showcase the powerful interpretability of structure-based methods through a case study on the Golgi apparatus, where we discover a decisive localization pattern α-helix from attention mechanisms, demonstrating the potential for bridging the gap with intuitive biological interpretability and paving the way for data-driven discoveries in cell biology.

1 INTRODUCTION

Understanding the subcellular localization of proteins is a fundamental question in cell biology, as a protein's function is often tightly coupled to its spatial context within the cell (Scott et al., 2005). Localization information is essential for elucidating molecular mechanisms such as signal transduction, metabolic regulation, and organelle-specific functions (Hung et al., 2017). It also provides a foundation for translational applications such as drug design (Hung et al., 2017; Rajendran et al., 2010). Recently, the data-driven AI approaches have emerged as a powerful paradigm for predicting whether or not a protein will be localized to a specific subcellular location. These methods substantially reduce the time and cost associated with traditional experimental techniques while holding promise for revealing novel biological patterns, thereby showcasing promising performance and attracting extensive research attention (Thumuluri et al., 2022; Stärk et al., 2021; Almagro Armenteros et al., 2017; Kobayashi et al., 2022; Elnaggar et al., 2021).

However, there remains a significant scarcity of high-quality datasets designed for this task. To the best of our knowledge, the only widely accepted dataset targeting this problem in the AI field is DeepLoc (Thumuluri et al., 2022; Almagro Armenteros et al., 2017), which contains the amino acid sequence information for each protein. DeepLoc has spurred the development of numerous sequence-based models for subcellular localization that infer localization solely from amino acid sequences. Nevertheless, several studies have shown that **spatial conformations** play a critical role in determining subcellular localization patterns. For example, the nuclear localization signals of transcription factor NF-κB are conditionally exposed only under specific structural conformations (Lusk et al., 2007). This demonstrates that the 3D structures of proteins, as dynamic regulatory elements, is the key to governing its subcellular localization.

To fully leverage protein structural data, recent research has developed structure-based protein representation models. Benefiting from the emergence of AlphaFold2 (Jumper et al., 2021), which offers reliable structural predictions for a vast number of proteins, the structure-based methods learn

representations directly from the spatial geometry of proteins. Such approaches have demonstrated impressive performance across a range of tasks, including protein classification (Jing et al., 2020; Zhang et al., 2022; Fan et al., 2022) and protein generation (Dauparas et al., 2022; Watson et al., 2023), showcasing their ability to capture complex structural patterns beyond what sequence alone can provide. These successful implementations underscore the substantial potential of incorporating structural information into subcellular localization prediction frameworks.

However, the existing subcellular localization datasets, such as DeepLoc, suffer from several limitations, which hinder the investigation of structure-based methods. Most notably, 1) they lack explicit protein 3D information, which is the key input to structure-based methods. Furthermore, 2) the current dataset typically uses coarse-grained compartment classifications, grouping subcellular areas into broad categories (*e.g.,* do not distinguish nuclear membrane and nucleoli in nucleus), which overlooks the unique localization characteristics and mechanisms associated with different organelles. Therefore, it leads to poor interpretability and great difficulty in discovering distinct patterns and underlying biological principles.

To address these limitations, we aim to construct a human protein subcellular localization dataset that can facilitate research on structure-based methods for localization prediction and enable the discovery of more specific and biologically relevant localization patterns. Specifically, we have two considerations for the dataset: 1) **Comprehensive 3D information**, which seeks to enhance the comprehensiveness of the dataset by recording detailed localization data from different databases and integrating 3D structural information of proteins, thereby bringing convenience and providing a unified evaluation benchmark for structure-based prediction models within the community; 2) **Fine-grained subcellular categorization**, which aims to incorporate finer-grained localization labels with annotations based on biological empirical evidence. As such, researchers are allowed to investigate protein localization patterns at a more detailed and functionally meaningful level.

To this end, we take the initiative of building a dataset called **CAPSUL** that simultaneously fulfills the two considerations. Specifically, to obtain the 3D information, we leverage AlphaFold2 to extract the Cartesian coordinates of the Cα (alpha carbon) and utilize the FoldSeek to derive corresponding 3Di structural tokens for each protein, promoting structure understanding such as backbone conformation, folding patterns, and local structure. Moreover, to obtain comprehensive subcellular localization labels, we cross-reference each protein with annotation data from both the UniProt (Consortium, 2019) and Human Protein Atlas (HPA) (Thul et al., 2017) databases. Building upon the categories in the existing dataset DeepLoc, we further refine the subcellular area space by introducing 20 aggregated subcellular compartments, carefully curated and validated by domain experts. We extend several state-of-the-art (SOTA) protein representation models to this downstream task and evaluate their performance on CAPSUL. To facilitate future research, we investigate several potential optimization strategies for structure-based model training and make innovative use of the attention mechanism to enhance the interpretability of protein subcellular localization patterns by integrating Transformer modules into existing models. Empirical results on CAPSUL validate the necessity of 3D information incorporation and the potential of leveraging structure-based methods for causal biology pattern discovery on the subcellular localization task.

In summary, the contributions of this paper are threefold:

- We represent the first systematic attempt to construct a human protein subcellular localization dataset with comprehensive 3D information, fine-grained categorization of cell compartments, and cross-referenced localization labels with experiment-level annotations.
- We evaluate several SOTA baseline models on our proposed dataset CAPSUL, highlighting the positive influence of incorporating protein structural information.
- We investigate various training strategies to facilitate future exploration and enhance the interpretability for subcellular localization tasks by introducing the attention mechanism.

2 Related Work

Sequence-based protein representation learning. Due to the relative ease of modeling protein amino acid sequences, early protein representation learning efforts typically relied solely on one-dimensional sequence inputs. Examples include models based on CNN, LSTM, or ResNet archi-

tectures (Shanehsazzadeh et al., 2020; Rao et al., 2019). Subsequently, Transformer-based models have demonstrated strong performance, especially after large-scale pretraining, achieving impressive results across a range of downstream tasks (Rives et al., 2019; Lin et al., 2022; Madani et al., 2023). In parallel, various self-supervised approaches have further enhanced the model's ability to capture meaningful features from protein sequences without a vast number of annotations (Rives et al., 2019; Lin et al., 2023; Elnaggar et al., 2021; Lu et al., 2020; He et al., 2021). However, in the subcellular localization task, which is known to be closely linked to protein structure, sequence-only models fall short of capturing the full complexity of protein features. As a result, incorporating 3D structural information has become increasingly recognized as essential for achieving richer and more comprehensive protein representations.

Structure-based protein representation learning. Efforts to model protein structures have been explored from multiple perspectives, including representations at the protein surface level, residue level, and atomic level. The protein language model also starts to consider structural information as input to enhance its understanding of proteins (Hayes et al., 2025). These approaches have achieved impressive results in tasks such as protein design, structure generation, and function prediction (Gligorijević et al., 2021; Gainza et al., 2020; Hermosilla et al., 2020; Hsu et al., 2022). Among them, models based on Graph Convolutional Network (GCN) have demonstrated consistently strong performance across various downstream tasks, highlighting their ability to effectively capture and interpret structural information (Fan et al., 2022; Jing et al., 2020; Zhang et al., 2022). However, most of these models require atomic or residue-level coordinate inputs, which are often missing from current benchmark datasets. To address this gap, we aim to construct a dataset specifically for the task of subcellular localization that incorporates 3D structural information, facilitating both the application and evaluation of structure-based models.

Subcellular localization dataset. Although many prestigious and task-specific protein benchmarks exist (Rao et al., 2019; Kryshtafovych et al., 2023), their lack of subcellular localization annotations makes them inapplicable on this downstream task. To the best of our knowledge, the only well-known dataset for subcellular localization originates from the training data used in DeepLoc (Thumuluri et al., 2022). Building on this, the PEER framework established a benchmark to evaluate baseline models on that dataset (Xu et al., 2022). However, the absence of 3D structural information makes it impossible to assess the performance of structure-based models that have already shown significant promise. To address this gap, we aim to reorganize and enrich the existing dataset by incorporating high-quality 3D structural information alongside fine-grained subcellular localization annotations. We further evaluate a range of representative baseline models on this updated dataset, with the goal of establishing a leading benchmark for subcellular localization prediction.

3 CAPSUL DATASET

To construct the CAPSUL dataset that offers 1) diverse and accessible 3D structural information, and 2) both detailed and aggregated subcellular localization annotations, we follow a multi-step curation process, as illustrated in Figure 1.

3.1 PROCESSING OF PROTEIN SEQUENCE AND STRUCTURE DATA

Collection and filter of protein data. We first retrieve all predicted human protein structures from the AlphaFold2 database (Jumper et al., 2021; Varadi et al., 2024), totaling 20,504 unique proteins. To ensure data quality and relevance, we filter this set by retaining only proteins marked as active in the UniProt database (Consortium, 2019), one of the most comprehensive and authoritative protein databases with well-documented annotations, resulting in a refined set of 20,401 proteins.

Removal of fragmented structure predictions. Among the refined set, AlphaFold2 typically adopts a sliding-window strategy to long protein sequences that segments the sequence with overlapping fragments to predict protein structure. To avoid inconsistencies of predicted coordinates that may arise during the stitching of these fragmented protein structures, we exclude such proteins from the dataset. After this step, we obtain 20,181 proteins of high quality and good consistency.

Extraction and preprocessing of protein features. We preserve the full PDB files for each protein, the original files downloaded from AlphaFold, including the positions of backbone atoms, side chains, and other relevant structural features essential for molecular modeling and analysis. The

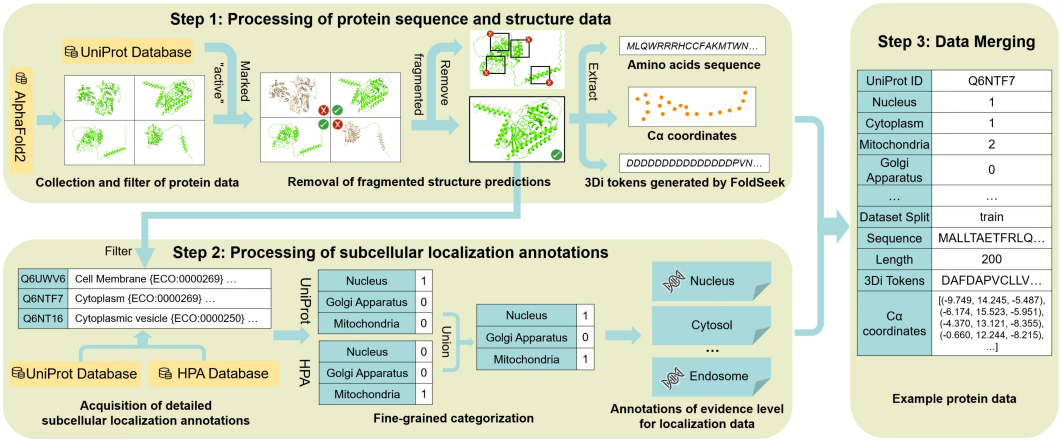

Figure 1: Procedures of CAPSUL dataset construction, including 3 key steps: Step 1 extracts and filters the sequence and structure data for each high-quality protein from AlphaFold2; Step 2 collects the annotations from UniProt and HPA for the resulting proteins in Step 1; Step 3 merges the structure data and the annotations for each protein, which consists of protein ID, localization annotations, amino acid sequence, sequence length, 3Di tokens, and Cα coordinates, *etc.*

coordinates of Cα atoms are extracted, which are important components for protein structure understanding. Furthermore, we employ the FoldSeek (Van Kempen et al., 2024) toolkit to tokenize the 3D structure of each amino acid. This provides a compact, informative structural representation that supports rapid, accurate modeling while reducing computational overhead, which has been empirically justified as effective and widely adopted in recent studies (Su et al., 2023).

Following the procedures above, we curate a dataset comprising 20,181 proteins, each labeled with amino acid sequence, Cα coordinates, and 3Di tokens sequence. For the next step, we append localization annotations to each protein.

3.2 Processing of Subcellular Localization Annotations

Acquisition of detailed subcellular localization annotations. Based on the obtained proteins above, we collect the corresponding detailed subcellular localization annotations for human proteins from both the UniProt and HPA databases. This detailed dataset provides high-resolution localization annotations on widely accepted subcellular compartments, which is vital to facilitate research into the specific localization patterns within distinct organelles.

Fine-grained categorization. After that, we aggregate the dataset by adopting a refined categorization approach. Specifically, we consolidate the subcellular locations into 20 distinct categories inspired by DeepLoc's and HPA's subcellular localization classification scheme (Thumuluri et al., 2022; Thul et al., 2017), which is a fine-grained framework compared with DeepLoc's ten-class categorization. Then, the sublocations of 20 categories are specified separately, so the various terminologies in different databases can align with 20 unified categorizations. The entire procedure was conducted in accordance with a well-established cell biology textbook (Alberts et al., 2022) and further verified by domain experts, with detailed categorization information available in *Supp.* A.

Annotations of evidence level for localization data. To fulfill the various research demands for the reliability of localization labels, we further extract and consolidate annotations on the experimental evidence level. Specifically, for UniProt, each subcellular localization annotation is accompanied by an evidence code indicating the source of the localization label. Among them, the localization supported by experimental evidence (marked with the term *ECO:0000269*) is labeled as 1, indicating experimental validation. For the localization with other forms of evidence (*e.g.,* non-traceable author statement evidence), the label 2 is assigned. The label 0 is assigned to the localizations without evidence annotations. Moreover, since HPA primarily relies on experimental data obtained through immunofluorescence and confocal microscopy (Thul et al., 2017), we assign label 1 to all annotated

Table 1: Comparisons between existing datasets and CAPSUL.

Dataset	Feartures		Categoization		Experimental Annotation
	Sequence	Structure	Aggregated	Detailed	
DeepLoc (Thumuluri et al., 2022)	✓	✗	✓	✗	✗
setHARD (Stärk et al., 2021)	✓	✗	✓	✗	✗
CAPSUL	✓	✓	✓	✓	✓

Table 2: Statistics of CAPSUL.

Number of Proteins	20,181	**Average Number of Annotations per Protein**	2.51	**Max Number of Annotations for Protein**	14	**Proportion of Experimental Annotations**	0.857
Number of Annotations on:							
Nucleus	7,590	Cytosol	5,386	Golgi Apparatus	1,881	Peroxisome	110
Nuclear Membrane	452	Cytoskeleton	2,119	Cell Membrane	5,777	Vesicle	2,863
Nucleoli	1,641	Centrosome	1,000	Endosome	687	Primary Cilium	983
Nucleoplasm	6,786	Mitochondria	1,768	Lipid Droplet	94	Secreted Proteins	2,087
Cytoplasm	6,613	Endoplasmic Reticulum	1,710	Lysosome/Vacuole	453	Sperm	652

localizations and label 0 to the localizations without evidence annotations. During the union of UniProt and HPA datasets, we prioritize annotations with experimental evidence when available.

3.3 Data Merging

After the separate processing of protein sequence and structure data, along with the subcellular localization annotations, we merge the data to include complete information. In Figure 1, we present a sample record in CAPSUL, which consists of protein ID, localization annotations, amino acid sequence, sequence length, 3Di tokens, and Cα coordinates, *etc.*

3.4 Dataset Analysis

In summary, we construct a unified dataset comprising 20,181 proteins, each annotated with 20 subcellular localization labels. Our dataset CAPSUL provides a more comprehensive coverage compared to DeepLoc (Thumuluri et al., 2022) and setHARD (Stärk et al., 2021) in terms of involved features, localization categorization, and experimental annotations, which is shown in Table 1. The dataset is randomly split into training, validation, and test sets in a 70%:15%:15% ratio for training and evaluation. We present a statistical analysis of numerical features of our dataset in Table 2.

To ensure the **high quality** of our constructed CAPSUL dataset, we have incorporated three safeguards[1]: 1) **Reliable data sources**: reliable protein structures predicted by AlphaFold2 were utilized in CAPSUL, with high accuracy, strong consistency, and incorporation of available experimental data as templates in its prediction process (Jumper et al., 2021); the localization labels source UniProt, a world-leading database with the most comprehensive protein annotations from multiple resources, and HPA, a human-specific protein database offering high-resolution and experiment-validated data. 2) **Strict validation and filtering**: we perform a series of validation and filtering steps on human proteins to exclude fragmented AlphaFold structures, which could introduce inconsistent coordinate information, and to remove proteins annotated as inactive in UniProt, thereby ensuring the reliability of subcellular localization annotations; 3) **Evidence-level support**: we incorporate annotations indicating whether experimental validation exists for the localization labels, thereby enhancing their credibility and catering to diverse research needs.

4 Experiments

4.1 Baseline Models

To study how existing methods perform on our proposed dataset, we evaluate 1) **DeepLoc 2.1** (Ødum et al., 2024), one of the most well-known tools dedicated to subcellular localization. It leverages the pre-trained protein language model ESM-1b (Rives et al., 2021) and provides predictions across ten subcellular compartments. Besides, we evaluate existing representative protein representation methods for the subcellular localization task, including sequence-based and structure-based methods.

[1]For a detailed analysis of the data reliability in the dataset, please refer to *Supp.* B.

Sequence-based models. Since existing sequence-based works are not specifically designed for subcellular tasks, we extend the widely adopted pre-trained protein language model 2) **ESM-2** (650M parameters) (Lin et al., 2022) and its latest iteration, 3) **ESM-C** (600M parameters) (ESM Team, 2024). We adopt the sequence encoder module from existing methods to obtain protein representation, and extend it with a localization classifier, as detailed in the following.

- *Sequence Encoder*. For each protein, we have its amino acid sequence represented as $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n) \in \mathbb{R}^{n\times 1}$, where s_i denotes the i-th residue and n is the length of the protein. We then apply the sequence encoder $f_{\text{seq}}(\cdot)$ of existing work to obtain contextual embeddings, $\boldsymbol{H} = f_{\text{seq}}(S)$, where $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n) \in \mathbb{R}^{n\times d}$, and $\boldsymbol{h}$ is the per-residue embeddings of dimension d. To obtain a fixed-length representation for the entire protein, we apply mean pooling and generate a global representation $\bar{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_i$, $\bar{\boldsymbol{h}} \in \mathbb{R}^d$.

- *Localization Classifier*. To predict subcellular localization, we leverage an MLP classifier $\phi(\cdot)$ on top of sequence encoder, *i.e.,* $\hat{\boldsymbol{y}} = \phi(\bar{\boldsymbol{h}})$, where $\hat{\boldsymbol{y}} \in \mathbb{R}^m$ is a multi-label prediction vector and m denotes the total number of predicted subcellular compartments.

Structure-based models. We consider 4) **CDConv** (Fan et al., 2022) and 5) **GearNet-Edge** (Zhang et al., 2022), two representative GCN baselines in protein representation task. We adopt the GCN-based structure encoder and extend it with an additional Transformer encoder to enhance interpretability. We also evaluate 6) **FoldSeek** (Van Kempen et al., 2024), which leverages a pre-trained structure tokenizer to encode the 3D structural information of each residue into a sequence of structure tokens. We also extend two novel methods, 7) **Graph Transformer** (Rampášek et al., 2022) and 8) **Graph Mamba** (Gu & Dao, 2023), to this task. The Graph Transformer employs attention mechanisms over graph-structured data, enabling the model to effectively capture both local and global dependencies among residues. Graph Mamba, on the other hand, incorporates selective state space models into graph learning, which facilitates long-range information propagation with improved efficiency. The outputs of the above models are then averaged and processed through a localization classifier for prediction.

- *Structure Encoder*. We represent a protein's 3D structure as a graph $G = (V, E)$, where each node $v_i \in V$ corresponds to the i-th residue (typically using the Cα atom position), and edges $(v_i, v_j) \in E$ are defined based on spatial or sequential adjacency. Each node v_i is initialized with a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$ including its positional information. Then we employ different graph encoders to capture higher-order topological relationships and produce updated representations $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n)$. The protein-level embedding is then obtained via global pooling $\bar{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_i$.

- *Localization Classifier*. We then obtain the final prediction $\hat{\boldsymbol{y}} = \phi\left(\bar{\boldsymbol{h}}\right)$, as described above.

Optimization. To optimize the models, we adopt the Binary Cross Entropy (BCE) loss, defined as $\mathcal{L}_{\text{BCE}} = -\frac{1}{m}\sum_{i=1}^{m}\left[\boldsymbol{y}_i \log(\hat{\boldsymbol{y}}_i) + (1-\boldsymbol{y}_i)\log(1-\hat{\boldsymbol{y}}_i)\right]$, where m is the number of classes, $\boldsymbol{y}_i \in \{0, 1\}$ is the label for class i, and $\hat{\boldsymbol{y}}_i \in (0, 1)$ is the predicted probability.

4.2 Benchmark Overall Results and Discussion

Given the class imbalance in each location (*i.e.*, the proportion of proteins localized to each subcellular compartment is often small), we consider the widely used evaluation metrics in this task: Precision, Recall, and F1-score (Jiang et al., 2021; Thumuluri et al., 2022). In addition, we utilize micro-averaged and macro-averaged F1-score to evaluate the overall performance across different categories. The overall performance[2] of all baselines on our proposed dataset is presented in Table 3, from which we have the following observations:

Large pre-training benefits sequence-based methods for subcellular location prediction. Among all sequence-based methods, ESM-C generally obtains higher F1-scores than ESM-2. We believe this is attributed to the extensive data and training compute used in the ESM-C pre-training, which facilitates a better representation of the protein's sequence features. Similar observations are also seen in (Hayes et al., 2025). Besides, this hypothesis can be further confirmed by the significantly inferior performance of ESM-C 600M[0], *i.e.,* without pre-training, than the pre-trained ESM-C. On the other hand, it is expected that DeepLoc yields inferior performance due to its overlook of

[2]The detailed results *w.r.t.* Precision and Recall, including other experimental results mentioned later in the main text, are provided in *Supp.* D.

Table 3: Overall performance of sequence-based and structure-based methods on CAPSUL.

Subcellular Locations	DeepLoc 2.1	ESM-2 650M	ESM-2 650M[f]	ESM-C 600M	ESM-C 600M[f]	ESM-C 600M[0]	FoldSeek	Graph Transformer	Graph Mamba	CDConv[t]	GearNet-Edge[t]
F1-score											
Nucleus	0.152	-	0.609	**0.649**	0.648	0.555	0.484	0.597	0.559	0.620	0.521
Nuclear Membrane	/	-	-	-	-	-	-	-	**0.037**	-	-
Nucleoli	/	-	-	0.091	0.039	0.024	-	**0.203**	0.168	0.147	0.121
Nucleoplasm	/	-	0.562	0.621	**0.623**	0.500	0.433	0.552	0.502	0.583	0.515
Cytoplasm	0.154	-	0.248	0.536	**0.551**	0.438	0.174	0.393	0.418	0.483	0.495
Cytosol	/	-	-	0.392	0.380	0.169	0.003	0.248	**0.426**	0.353	0.385
Cytoskeleton	/	-	0.006	0.251	0.205	0.048	0.070	0.042	**0.270**	0.135	0.228
Centrosome	/	-	-	0.014	-	-	-	-	**0.181**	-	0.127
Mitochondria	0.120	-	0.317	**0.562**	0.544	0.099	-	0.475	0.341	0.476	0.318
Endoplasmic Reticulum	0.121	-	-	**0.351**	0.333	0.059	-	0.184	0.059	0.292	0.279
Golgi Apparatus	0.061	-	-	0.099	0.027	-	-	0.041	**0.185**	0.073	0.026
Cell Membrane	0.142	-	0.555	0.631	**0.648**	0.372	0.343	0.547	0.540	0.562	0.556
Endosome	/	-	-	0.018	-	-	-	-	**0.100**	-	0.067
Lipid Droplet	/	-	-	-	-	-	-	-	-	-	-
Lysosome/Vacuole	**0.118**	-	-	-	-	-	-	-	-	-	0.073
Peroxisome	**0.131**	-	-	-	-	-	-	-	-	-	-
Vesicle	/	-	-	0.009	-	0.005	-	0.044	**0.135**	0.027	0.068
Primary Cilium	/	-	-	**0.164**	0.112	-	-	0.012	0.088	-	0.147
Secreted Proteins	0.191	-	0.713	**0.826**	0.797	0.433	0.328	0.705	0.557	0.767	0.687
Sperm	/	-	-	0.052	0.070	-	-	0.018	**0.130**	-	0.086
Micro Avg F1-score	/	-	0.375	**0.495**	0.492	0.338	0.248	0.410	0.411	0.452	0.417
Macro Avg F1-score	/	-	0.150	**0.263**	0.249	0.135	0.092	0.203	0.235	0.226	0.235
Micro Avg Precision	/	-	0.647	0.690	**0.693**	0.598	0.605	0.637	0.414	0.632	0.546
Micro Avg Recall	/	-	0.264	**0.386**	0.382	0.236	0.156	0.302	0.408	0.352	0.337

[f]We finetune the pre-trained protein language model. [t]The original MLP is replaced by Transformer layers. [0]The parameters of ESM-C are initialized randomly. "/" indicates that DeepLoc 2.1 does not support prediction for that location, and therefore, average metrics are not considered in this case. "–" indicates that no prediction is made for that location. **Bold** value indicates the best results.

Table 4: Ablation study of CDConv and GearNet-Edge to randomly sample Cα coordinates.

	CDConv[t] (random Cα coordinates)	CDConv[t]	GearNet-Edge[t] (random Cα coordinates)	GearNet-Edge[t]
Micro Avg F1-score	0.329	**0.452**	0.348	**0.417**
Micro Avg Precision	0.586	**0.632**	0.450	**0.546**
Micro Avg Recall	0.229	**0.352**	0.283	**0.337**

[t]The original MLP is replaced by Transformer layers. **Bold** value indicates the better result for each baseline.

the fine-grained categorization during pre-training, which may result in its inability to sufficiently differentiate the representations of proteins in multi-label classification tasks (Hong et al., 2023). This further validates the necessity of detailed categorizations of subcellular locations in CAPSUL.

The 3D structure is essential for subcellular localization task. Despite that structure-based methods slightly fall behind the pre-trained ESM-C, both CDConv and GearNet-Edge outperform the ESM-C 600M[0] in most cases. Also, a group of ablation studies is conducted on CDConv and GearNet-Edge, with coordinates randomly sampled from each protein's spatial range. As shown in Table 4, randomly sampling the input of protein 3D structural data leads to a significant drop in model performance. These two results validate that structural information plays a decisive role in determining subcellular localization. Besides, CDConv demonstrates the strongest overall performance among the structure-based models, justifying the effectiveness of relative distance and the dynamic radius for convolution. Nevertheless, the inferior performance of FoldSeek may be due to the lack of sequence information and its coarse tokenization of structural information.

The models generally demonstrate better performance on subcellular locations with larger localization sample sizes. For classes with a large number of localization samples (*e.g.*, nucleus), most models tend to demonstrate relatively strong predictive performance. In contrast, for underrepresented classes (*e.g.*, lipid droplet), the prediction performance is generally poor, with some classes even failing to produce any correctly identified proteins. This is a common outcome in imbalanced multi-label classification tasks, as the standard BCE loss tends to neglect fewer-number labels. Additionally, potential conflicts among multiple optimization targets may further exacerbate this issue. To address these challenges, we conduct in-depth analysis in Section 4.3.1 and 4.3.2, exploring strategies such as reweighting and single-label classification to mitigate the effects of class imbalance and task conflict.

Structure-based models showcase their potential to capture non-trivial patterns for subcellular locations with few samples. Graph Mamba and GearNet-Edge tend to perform better on certain classes with smaller localization sample sizes. We believe that this is because of the relational message passing layer adopted in them, which uniquely models different spatial interactions among residues. This demonstrates that structure-based models showcase potential to identify spe-

Table 5: Performance of ESM-C 600M, CDConv, and GearNet-Edge with reweighting scheme.

Subcellular Locations	ESM-C 600M	CDConv†	GearNet-Edge†	Subcellular Locations	ESM-C 600M	CDConv†	GearNet-Edge†
F1-score				**F1-score**			
Nucleus	0.630	**0.625**	**0.618**	Endosome	-	**0.114**	**0.150**
Nuclear Membrane	-	**0.062**	**0.058**	Lipid Droplet	**0.235**	**0.023**	**0.111**
Nucleoli	-	**0.188**	**0.224**	Lysosome/Vacuole	-	**0.175**	**0.111**
Nucleoplasm	0.576	**0.607**	**0.574**	Peroxisome	**0.190**	**0.072**	**0.108**
Cytoplasm	0.500	**0.582**	**0.544**	Vesicle	-	**0.288**	**0.281**
Cytosol	**0.133**	**0.495**	**0.484**	Primary Cilium	0.024	**0.167**	**0.176**
Cytoskeleton	0.083	**0.292**	**0.294**	Secreted Proteins	0.778	0.564	0.614
Centrosome	-	**0.160**	**0.175**	Sperm	-	**0.120**	**0.125**
Mitochondria	0.481	0.297	0.313				
Endoplasmic Reticulum	-	**0.308**	**0.345**				
Golgi Apparatus	-	**0.246**	**0.238**	**Micro Avg F1-score**	0.429	0.381	**0.453**
Cell Membrane	0.566	0.560	0.536	**Macro Avg F1-score**	0.210	**0.297**	**0.304**

†The original MLP is replaced by Transformer layers. "–" indicates that no prediction is made for that location. **Bold** value indicates that it improves compared with the result without reweighting.

Table 6: Performance of ESM-C 600M, CDConv, and GearNet-Edge with single-label classification.

Subcellular Locations	ESM-C 600M	CDConv†	GearNet-Edge†	Subcellular Locations	ESM-C 600M	CDConv†	GearNet-Edge†
F1-score				**F1-score**			
Nuclear Membrane	-	**0.052**	**0.042**	Lysosome/Vacuole	**0.115**	-	**0.162**
Nucleoli	**0.267**	**0.151**	**0.228**	Peroxisome	**0.054**	-	**0.023**
Centrosome	**0.184**	**0.089**	**0.167**	Vesicle	**0.068**	**0.230**	**0.268**
Golgi Apparatus	**0.280**	**0.114**	**0.210**	Primary Cilium	**0.253**	**0.097**	**0.171**
Endosome	**0.167**	**0.049**	**0.126**	Sperm	**0.159**	**0.068**	**0.117**
Lipid Droplet	**0.021**	-	**0.051**				

†The original MLP is replaced by Transformer layers. "–" indicates that no prediction is made for that location. **Bold** value indicates that it improves compared with the result from multi-label classification.

cific structural features that are indicative of localization to a particular organelle, thus achieving a notably good performance. Further investigation on the patterns with intuitive biological interpretability captured by the structure-based model can be found in Section 4.3.3.

4.3 In-depth Analysis

4.3.1 Protein Imbalance Mitigation via Reweighting

Reweighting Schemes. In this task, for each subcellular location, the number of positive samples (*i.e.*, proteins localized to that compartment) is substantially smaller than the number of negative samples (*i.e.*, proteins not localized to that compartment). Reweighting is a widely used strategy to address class imbalance by reducing the bias toward majority classes. Inspired by previous work on class-level reweighting, we evaluate three reweighting schemes. 1) Inverse frequency reweighting (Cao et al., 2019), *i.e.,* $w_c = \frac{1}{f_c}$. 2) Log-inverse frequency reweighting (Cui et al., 2019), *i.e.,* $w_c = \frac{1}{\log(1+f_c)}$. 3) Focal loss (Lin et al., 2017), which is defined as

$$\mathcal{L}_c = -w_c \cdot \sum_i \left[y_{ic} \cdot (1-\hat{y}_{ic})^\gamma \log(\hat{y}_{ic}) + (1-y_{ic}) \cdot \hat{y}_{ic}^\gamma \log(1-\hat{y}_{ic}) \right],$$

where f_c is the frequency of positive samples in class c, w_c is the computed class-specific weight, $y_{ic} \in \{0,1\}$ denotes the ground truth label for sample i and class c, $\hat{y}_{ic} \in (0,1)$ is the predicted probability, and γ is the focusing parameter. It deserves attention that the w_c in the focal loss strategy is chosen from either inverse or log-inverse frequency weight.

Results. We apply the three reweighting schemes on three competitive models (ESM-C, CDConv, and GearNet-Edge) and report the best results for each model in Table 5. From the results, we observe that the two structure-based baseline models exhibit substantial improvements under reweighting strategies, especially for the higher Precision across underrepresented categories. In particular, CDConv and GearNet-Edge successfully identify positive instances for every class. These findings highlight that reweighting can significantly enhance model performance on minority classes, especially for structure-based models.

4.3.2 Single-label Classification

To explore how different methods perform on each subcellular location respectively, we adopt the single-label setting, aiming to mitigate the potential conflict between optimization across different

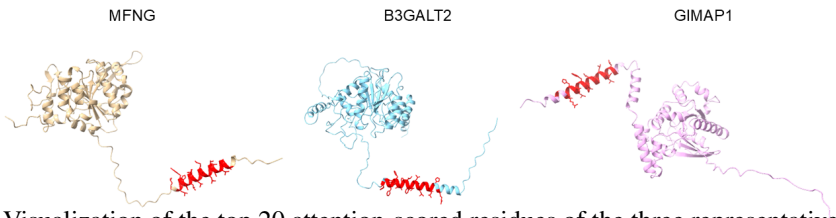

Figure 2: Visualization of the top 20 attention-scored residues of the three representative proteins.

classes. In this setting, we train separate binary classifiers for each subcellular localization category with ESM-C, CDConv, and GearNet-Edge. We apply this single-label prediction framework specifically to those subcellular localization classes where the F1-score of at least one of the models (ESM-C, CDConv, or GearNet-Edge) is lower than 0.1. Our goal is to shift the model's attention toward underrepresented classes and improve the predictive influence of positive samples.

From the results in Table 6, we observe 1) notable improvements in the prediction performance of previously underperforming classes, particularly for GearNet-Edge. However, 2) ESM-C and CDConv still fail to generate any predictions for a few categories, primarily due to the extremely low proportion of positive samples (ranging from 0.5% to 3%). Given that such severe class imbalance is a common challenge in subcellular localization tasks, we consider the single-label prediction strategy a promising and practical solution. Moreover, this approach lays the groundwork for future research focused on identifying localization patterns specific to individual subcellular compartments.

4.3.3 Biological Interpretability

We analyze a CDConv model on Golgi apparatus prediction with an exceptional precision of 100%. Specifically, with our novel attempt of the Transformer module extended to the GCN-based models, we identify and visualize the tokens (*i.e.*, residues) that receive the 20 highest attention weights in Figure 2, offering insights into which structure the model considers most decisive for subcellular localization. We find that the model consistently highlights similar α-helix spatial conformation, such as residues 8-27 of MFNG, residues 24-45 of B3GALT2, and residues 273-292 at the C-terminus of GIMAP1. Remarkably, these findings show strong concordance with prior experimental evidence (Paulson & Colley, 1989; Linstedt et al., 1995). It is highlighted that despite significant sequence divergence, the model specifically focuses on α-helix transmembrane domains (20-30 amino acids in length) that maintain consistent topological orientations across all targets. Recent studies have demonstrated that the topological conformation of transmembrane domains can influence Golgi localization by regulating electrostatic potential gradients in transmembrane regions and lipid membrane anchoring efficiency (Cosson et al., 2013; Hanulova & Weiss, 2012; Bian et al., 2024). This evidence not only confirms the model's capability for structural pattern recognition beyond sequence similarity but also provides theoretical support for its structural identification mechanisms.

5 Conclusion and Future Work

We pointed out the crucial importance of constructing a subcellular localization benchmark with protein 3D information to facilitate the investigation of structure-based models for the subcellular localization task. To achieve this, we constructed a benchmark called CAPSUL that contains comprehensive structural information and fine-grained annotations of 20 categories of subcellular compartments with biological experiment evidence labels. Based on CAPSUL, we evaluated SOTA sequence-based and structure-based models as well as their feasible optimization strategies, demonstrating the effectiveness of incorporating protein structural information. Moreover, a case study on Golgi apparatus validates the biology-aligned interpretability of structure-based models trained on a specific fine-grained subcellular location, supported by CAPSUL. This work proposes a comprehensive human protein benchmark with 3D information and fine-grained annotations for subcellular localization, Based on CAPSUL, we highlight several research directions that are worth future exploration: 1) To fully leverage structural information, aligning or disentangling the understanding across different dimensions (*i.e.,* amino acid sequence, Cα, and 3Di) specifically for subcellular localization is a promising direction. 2) Causal discovery on the relationship between 3D structure and subcellular localization is worthwhile to be explored on CAPSUL, with the goal of establishing direct links to underlying biological principles.

Ethics Statement

This research presents a dataset and benchmark for protein subcellular localization prediction using AI methods. We confirm that our work raises no ethical concerns as it involves only the analysis of publicly available protein data, with no human subjects, animal experiments, or biological interventions. We have fully considered the potential societal impacts and do not foresee any direct, immediate, or negative consequences. We are committed to the ethical dissemination of our findings and encourage their responsible use.

Reproducibility Statement

All the results in this work are reproducible. The access to the necessary code and complete dataset can be found in *Supp.* H. We discuss the experimental details in *Supp.* C, including implementation details such as the hyperparameters chosen for each experiment, to help reproduce our results. Additionally, further experimental results, detailed dataset interpretations, and usage guidelines are provided in *Supp.* E, F, and G to facilitate better understanding and utilization of our dataset.

References

Bruce Alberts, Rebecca Heald, Alexander Johnson, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell: Seventh edition*. Norton and Company, 2022.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Claudie Bian, Anna Marchetti, Marco Dias, Jackie Perrin, and Pierre Cosson. Short transmembrane domains target type ii proteins to the golgi apparatus and type i proteins to the endoplasmic reticulum. *Journal of Cell Science*, 137(15), 2024.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47 (D1):D506–D515, 2019.

Pierre Cosson, Jackie Perrin, and Juan S Bonifacino. Anchors aweigh: protein localization and transport mediated by transmembrane domains. *Trends in cell biology*, 23(10):511–517, 2013.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7112–7127, 2021.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2022.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael M Bronstein, and Bruno E Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Maria Hanulova and Matthias Weiss. Membrane-mediated interactions–a physico-chemical basis for protein sorting. *Molecular Membrane Biology*, 29(5):177–185, 2012.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, et al. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*, 2021.

Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *arXiv preprint arXiv:2007.06252*, 2020.

Guan Zhe Hong, Yin Cui, Ariel Fuxman, Stanley H Chan, and Enming Luo. Towards understanding the effect of pretraining label granularity. *arXiv preprint arXiv:2303.16887*, 2023.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Victoria Hung, Stephanie S Lam, Namrata D Udeshi, Tanya Svinkina, Gaelen Guzman, Vamsi K Mootha, Steven A Carr, and Alice Y Ting. Proteomic mapping of cytosol-facing outer mitochondrial and er membranes in living human cells by proximity biotinylation. *elife*, 6:e24463, 2017.

Yuexu Jiang, Duolin Wang, Yifu Yao, Holger Eubel, Patrick Künzler, Ian Max Møller, and Dong Xu. Mulocdeep: a deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. *Computational and structural biotechnology journal*, 19:4825–4839, 2021.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Hirofumi Kobayashi, Keith C Cheveralls, Manuel D Leonetti, and Loic A Royer. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nature methods*, 19(8):995–1003, 2022.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xv. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1539–1549, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

AD Linstedt, M Foguet, M Renz, HP Seelig, BS Glick, and HP Hauri. A c-terminally-anchored golgi protein is inserted into the endoplasmic reticulum and then transported to the golgi apparatus. *Proceedings of the National Academy of Sciences*, 92(11):5102–5105, 1995.

Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv*, pp. 2020–09, 2020.

C Patrick Lusk, Günter Blobel, and Megan C King. Highway to the inner nuclear membrane: rules for the road. *Nature Reviews Molecular Cell Biology*, 8(5):414–420, 2007.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41 (8):1099–1106, 2023.

Marius Thrane Ødum, Felix Teufel, Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. Deeploc 2.1: multi-label membrane protein type prediction using protein language models. *Nucleic Acids Research*, 52(W1):W215–W220, 2024.

James C Paulson and Karen J Colley. Glycosyltransferases: structure, localization, and control of cell type-specific glycosylation. *Journal of Biological Chemistry*, 264(30):17615–17618, 1989.

Lawrence Rajendran, Hans-Joachim Knölker, and Kai Simons. Subcellular targeting strategies for drug design and delivery. *Nature reviews Drug discovery*, 9(1):29–42, 2010.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Michelle S Scott, Sara J Calafell, David Y Thomas, and Michael T Hallett. Refining protein subcellular localization. *PLoS computational biology*, 1(6):e66, 2005.

Amir Shanehsazzadeh, David Belanger, and David Dohan. Is transfer learning necessary for protein landscape prediction? *arXiv preprint arXiv:2011.03443*, 2020.

Hannes Stärk, Christian Dallago, Michael Heinzinger, and Burkhard Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 2021.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pp. 2023–10, 2023.

Peter J Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M Breckels, et al. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, 2017.

Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research*, 50(W1):W228–W234, 2022.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.

Supplementary Material for CAPSUL: A Comprehensive Human Protein Benchmark for Subcellular Localization

A Dataset Construction

A.1 Subcellular Location Categorization and Terminology Mapping

To facilitate model classification, we first categorize the detailed subcellular localizations of proteins. Existing datasets often use coarse-grained classifications (*e.g.,* DeepLoc categorizes subcellular locations into 10 broad classes). However, since each subcellular compartment typically follows distinct localization patterns, such coarse categorizations can hinder the model's ability to capture consistent intra-class features, ultimately leading to reduced prediction accuracy. Moreover, coarse-grained classification also hinders researchers from exploring localization mechanisms specific to finer subcellular compartments. Inspired by the subcellular location categories in HPA and DeepLoc, we propose a finer-grained classification scheme consisting of 20 subcellular categories. Notably, "Nucleus" and "Cytoplasm" categories serve as umbrella terms for several finer locations to ensure compatibility with DeepLoc during evaluation.

When aligning protein localization annotations from the UniProt and HPA databases to our refined categorization, we observe inconsistencies in terminology (*e.g.,* "Cell Membrane" in UniProt versus "Plasma Membrane" in HPA). To resolve such discrepancies, we refer to the prestigious textbook Molecular Biology of the Cell (7th Edition) (Alberts et al., 2022) and create a unified mapping, as shown in Table 7, which allows for consistent categorization across the two databases.

Domain experts were extensively engaged to ensure and validate the accuracy of the classification standards and data alignment procedures. We invited cell biologists from several prestigious universities and research institutes to review and revise the dataset, which ensures that CAPSUL is firmly grounded in cell biology. All of them have over eight years of research experience in their field. They are rigorously involved throughout the entire process, including 1) curating authoritative datasets, 2) determining primary subcellular localizations, and 3) validating the biological plausibility of localization assignments.

Through the above processes, we have established a fine-grained subcellular localization classification standard and successfully unified annotations from multiple databases under a unified labeling framework.

A.2 Dataset Splits

To construct separate datasets for training, validating, and testing, we randomly split the original dataset into three subsets in a 70%: 15%: 15% ratio. The partitioning of different protein data used in our experiments is also available in the CAPSUL dataset. The number of labels for each subcellular location in three subsets is shown in Table 8. Although the data is randomly assigned to different subsets, we have verified the distribution characteristics among classes to maintain a similar proportional relationship, ensuring balance and representativeness across the subsets.

B Dataset Reliability

In Section 3, we provide a detailed description of the data preprocessing procedures implemented to ensure the high quality of CAPSUL. Here, we would like to emphasize that the data sources themselves are highly reliable. Specifically, the protein-related data used in this study were primarily obtained from the following databases:

AlphaFold. AlphaFold provides protein structural data in CAPSUL. 1) AlphaFold has already **incorporated experimentally resolved structures of proteins as templates** during its prediction process (Jumper et al., 2021). AlphaFold explicitly describes how its pipeline automatically searches the PDB for experimentally resolved structures, selecting up to four structural templates, and maps atom coordinates from those templates to the target sequence during inference. These coordinates are used as template inputs alongside MSA-based evolutionary information, enabling AlphaFold to

Table 7: Categorization of CAPSUL and terminology mapping between HPA and Uniprot.

20 fine-grained categories	HPA	UniProt
Nucleus		
Nuclear Membrane	Nuclear membrane	Nucleus membrane, Nucleus envelope, Nucleus inner membrane, Nucleus outer membrane
Nucleoli	Nucleoli, Nucleoli fibrillar center, Nucleoli rim	Nucleolus
Nucleoplasm	Kinetochore, Mitotic chromosome, Nuclear bodies, Nuclear speckles, Nucleoplasm	Nucleus matrix, Nucleus lamina, Chromosome, Nucleus speckle
Cytoplasm		
Cytosol	Aggresome, Cytoplasmic bodies, Cytosol, Rods Rings	Cytosol
Cytoskeleton	Actin filaments, Cleavage furrow, Focal adhesion sites, Cytokinetic bridge, Microtubule ends, Microtubules, Midbody, Midbody ring, Mitotic spindle, Intermediate filaments	Cytoskeleton
Centrosome	Centriolar satellite, Centrosome	Centrosome
Mitochondria	Mitochondria	Mitochondrion, Mitochondrion envelop, Mitochondrion inner membrane, Mitochondrion outer membrane, Mitochondrion membrane, Mitochondrion matrix, Mitochondrion intermembrane space
Endoplasmic Reticulum	Endoplasmic reticulum	Endoplasmic reticulum, Endoplasmic reticulum membrane, Endoplasmic reticulum lumen, Microsome, Rough endoplasmic reticulum, Smooth endoplasmic reticulum, Sarcoplasmic reticulum
Golgi Apparatus	Golgi apparatus	Golgi apparatus, Golgi apparatus membrane, Golgi apparatus lumen
Cell Membrane	Cell Junctions, Plasma membrane	Cell membrane, Apical cell membrane, Apicolateral cell membrane, Basal cell membrane, Basolateral cell membrane, Lateral cell membrane, Cell projection
Endosome	Endosomes	Endosome
Lipid Droplet	Lipid droplets	Lipid droplet
Lysosome/Vacuole	Lysosomes	Lysosome, Vacuole, Vacuole lumen, Vacuole membrane, Lysosome lumen, Lysosome membrane
Peroxisome	Peroxisomes	Peroxisome, Peroxisome matrix, Peroxisome membrane
Vesicle	Vesicles	Vesicle
Primary Cilium	Basal body, Primary cilium, Primary cilium tip, Primary cilium transition zone	Cilium
Secreted Proteins	Secreted Proteins	Secreted
Sperm	Acrosome, Annulus, Calyx, Connecting piece, End piece, Equatorial segment, Flagellar centriole, Mid piece, Perinuclear theca, Principal piece	Acrosome, Calyx, Perinuclear theca

Table 8: Label counts for training, validation, and test set of CAPSUL.

Subcellular Locations	Counts			
	Training Set	Validation Set	Test Set	**Sum**
Nucleus	5,312	1,128	1,150	7,590
Nuclear Membrane	313	63	76	452
Nucleoli	1,143	249	249	1,641
Nucleoplasm	4,751	1,007	1,028	6,786
Cytoplasm	4,652	984	977	6,613
Cytosol	3,787	811	788	5,386
Cytoskeleton	1,499	302	318	2,119
Centrosome	713	140	147	1,000
Mitochondria	1,247	259	262	1,768
Endoplasmic Reticulum	1,146	275	289	1,710
Golgi Apparatus	1,323	271	287	1,811
Cell Membrane	4,022	863	892	5,777
Endosome	466	113	108	687
Lipid Droplet	63	16	15	94
Lysosome/Vacuole	313	65	75	453
Peroxisome	71	20	19	110
Vesicle	2,019	404	440	2,863
Primary Cilium	699	123	161	983
Secreted Proteins	1,477	317	293	2,087
Sperm	444	99	109	652

leverage high-quality experimental structural data in its predictions. 2) AlphaFold-predicted structures have been demonstrated to achieve exceptionally **high accuracy**, competitive with experimental data. AlphaFold was entered for CASP14, and shows that it achieves accuracy competitive with experiment in a majority of cases. Specifically, the median backbone accuracy of its predictions is 0.96 Å r.m.s.d.$_{95}$ (Cα root-mean-square deviation at 95% residue coverage), which is often within the margin of error of experimental structures (Jumper et al., 2021). 3) AlphaFold provides full-length protein structures containing complete structural information, which minimizes the potential negative influence of structural variability caused by different versions of experimental protein data. This choice allows us to maintain **a high level of consistency** across the CAPSUL dataset.

UniProt. UniProt provides protein localization annotation and evidence-level annotations in CAPSUL. UniProt serves as one of the most authoritative and widely used protein knowledge bases, integrating sequence, functional, and localization information across a broad spectrum of species. In particular, the manually curated Swiss-Prot section is recognized for its rigorous curation standards, where annotations (including subcellular localization annotations) are derived from authoritative experimental studies and peer-reviewed literature, complemented by computational analyses and homology-based inferences. Each localization entry is systematically annotated with evidence codes that explicitly denote whether the information originates from direct experimental validation, literature reports, or computational prediction, thereby providing transparency and traceability of the data source. This evidence-based framework ensures that localization annotations are not only comprehensive but also of consistently high quality.

Human Protein Atlas (HPA). HPA provides protein localization annotation and subcellular categories reference in CAPSUL. HPA provides a unique and experimentally grounded resource for human protein subcellular localization. Its Subcellular Atlas is built upon systematic immunofluorescence imaging combined with antibody-based profiling in multiple well-characterized human cell lines. This approach allows direct visualization of protein distribution within distinct subcellular compartments, thereby offering cell-type-specific and high-resolution localization evidence. These measures substantially reduce the likelihood of false annotations and provide users with a clear indication of annotation confidence.

C Experiment Details

C.1 Implementation Details

The experiments were performed utilizing NVIDIA RTX 3090, A40 and A100 GPUs. We employ an early stopping strategy to mitigate overfitting with a tolerance of 5 epochs. Hyperparameters

such as learning rate, number of epochs, and batch size are explored separately for each model type, considering their distinct architectures.

C.2 Description of Graph Encoder

Within the structure-based baseline models, the graph encoders vary in their approaches to processing the input feature vectors: A **GCN** updates node representations via neighborhood aggregation, *i.e.,* $\boldsymbol{m}_i^{(0)} = \boldsymbol{x}_i$, $\boldsymbol{m}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \boldsymbol{W}^{(l)}\boldsymbol{m}_j^{(l)} + \boldsymbol{b}^{(l)}\right)$. $\boldsymbol{m}_i^{(l)}$ is the representation of node i at layer l, $\mathcal{N}(i)$ denotes the neighbors of node i, $W^{(l)}$ and $\boldsymbol{b}^{(l)}$ are trainable weights and bias, and σ is a non-linear activation function (*e.g.*, ReLU). After L layers of graph convolution, we obtain the final node representations $\{\boldsymbol{m}_i^{(L)}\}_{i=1}^n$. To enhance interpretability and capture global interactions among residues, we replace the traditional average pooling with a Transformer encoder $\mathcal{T}(\cdot)$ to obtain the residue representation, *i.e.,* $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n) = \mathcal{T}\left(\{\boldsymbol{m}_i^{(L)}\}_{i=1}^n\right)$, where $\boldsymbol{h} \in \mathbb{R}^d$. Similarly, **Graph Transformer** and **Graph Mamba** substitute the convolution-based encoder with their respective architectures, while adhering to the same overall procedure to obtain the global protein representation.

C.3 Hyperparameter Settings

For all the experiments, we choose the best hyperparameters according to the best micro F1-score on the test set.

For the main experiment, the best hyperparameter setting for each model is as follows: 1) **ESM-2 (650M)**, the MLP hidden layers are set to (512,256), and learning rate to 1×10^{-4}. 2) **ESM-C (600M)**, the MLP hidden layers are set to (512,256) (to (512) when finetuning), and learning rate to 5×10^{-4}. 3) **FoldSeek**, the embedding dimensions are set to 256, transformer layers to 2, transformer heads to 4, and learning rate to 1×10^{-4}. 4) **CDConv**, the kernel channels are set to 24, feature channels to (256,512), geometric radius to 4.0, sequential kernel size to 5, transformer layers to 3, transformer heads to 2, and learning rate to 5×10^{-4}. 5) **GearNet-Edge**, the convolution hidden dimensions are set to (512,512,512), transformer layers to 2, transformer heads to 2, and learning rate to 1×10^{-5}. 6) **Graph Transformer**, the transformer layers are set to 10, node dimensions set to 256, positional embedding dimension set to 8, and learning rate set to 5×10^{-5}. 7) **Graph Mamba**, the Mamba layers are set to 5, node dimensions set to 256, and learning rate set to 1×10^{-4}.

For the reweighting strategy, we inherit the optimal hyperparameter settings for ESM-C (600M), CDConv, and GearNet-Edge mentioned above. The best reweighting scheme for each model is as follows: 1) **ESM-C (600M)**, focal loss with α set to the weights of log-inverse frequency, and γ set to 1.0. 2) **CDConv**, focal loss with α set to the weights of log-inverse frequency, and γ set to 3.0. 3) **GearNet-Edge**, inverse frequency reweighting.

For the single-label classification strategy, we inherit the optimal hyperparameter settings for ESM-C (600M), CDConv, and GearNet-Edge mentioned above. To address class imbalance, we undersample the negative class to achieve a 1:3 positive-to-negative sample ratio for ESM-C (600M), and a 1:1 positive-to-negative sample ratio for CDConv.

D Detailed Baseline Results

Detailed experimental results of main experiments, reweighting strategy, and single-label classification strategy are provided in Tables 9, 10, and 11, respectively. They include evaluation metrics of precision, recall, and F1-score.

E Ablation Study

Although it has been recognized in the biological community that many patterns of subcellular localization cannot be fully captured by simple sequence information, we aim to investigate the potential benefits of incorporating protein structural information as input for prediction. Therefore,

Table 9: Detailed performance of sequence-based and structure-based methods on CAPSUL.

Subcellular	DeepLoc 2.1			ESM-2 650M			ESM-2 650M[f]		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.675	0.086	0.152	-	-	-	0.633	0.586	0.609
Nuclear Membrane	/	/	/	-	-	-	-	-	-
Nucleoli	/	/	/	-	-	-	-	-	-
Nucleoplasm	/	/	/	-	-	-	0.592	0.535	0.562
Cytoplasm	0.510	0.100	0.167	-	-	-	0.598	0.157	0.248
Cytosol	/	/	/	-	-	-	-	-	-
Cytoskeleton	/	/	/	-	-	-	0.200	0.003	0.006
Centrosome	/	/	/	-	-	-	-	-	-
Mitochondria	0.799	0.065	0.120	-	-	-	0.850	0.195	0.317
Endoplasmic Reticulum	0.581	0.067	0.121	-	-	-	-	-	-
Golgi Apparatus	0.594	0.032	0.061	-	-	-	-	-	-
Cell Membrane	0.740	0.078	0.142	-	-	-	0.722	0.451	0.555
Endosome	/	/	/	-	-	-	-	-	-
Lipid Droplet	/	/	/	-	-	-	-	-	-
Lysosome/Vacuole	0.198	0.084	0.118	-	-	-	-	-	-
Peroxisome	0.667	0.073	0.131	-	-	-	-	-	-
Vesicle	/	/	/	-	-	-	-	-	-
Primary Cilium	/	/	/	-	-	-	-	-	-
Secreted Proteins	0.773	0.109	0.191	-	-	-	0.742	0.686	0.713
Sperm	/	/	/	-	-	-	-	-	-
Micro Avg	/	/	/	-	-	-	0.647	0.264	0.375
Macro Avg	/	/	/	-	-	-	0.217	0.131	0.150
Subcellular	ESM-C 600M			ESM-C 600M[f]			ESM-C 600M[0]		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.694	0.609	0.649	0.708	0.597	0.648	0.626	0.498	0.555
Nuclear Membrane	-	-	-	-	-	-	-	-	-
Nucleoli	0.800	0.048	0.091	1.000	0.020	0.039	1.000	0.012	0.024
Nucleoplasm	0.679	0.573	0.621	0.686	0.570	0.623	0.620	0.418	0.500
Cytoplasm	0.611	0.477	0.536	0.614	0.499	0.551	0.507	0.385	0.438
Cytosol	0.541	0.307	0.392	0.567	0.286	0.380	0.456	0.104	0.169
Cytoskeleton	0.681	0.154	0.251	0.629	0.123	0.205	0.471	0.025	0.048
Centrosome	1.000	0.007	0.014	-	-	-	-	-	-
Mitochondria	0.865	0.416	0.562	0.903	0.389	0.544	0.667	0.053	0.099
Endoplasmic Reticulum	0.687	0.235	0.351	0.674	0.221	0.333	0.500	0.031	0.059
Golgi Apparatus	0.938	0.052	0.099	1.000	0.014	0.027	-	-	-
Cell Membrane	0.777	0.531	0.631	0.753	0.568	0.648	0.786	0.243	0.372
Endosome	1.000	0.009	0.018	-	-	-	-	-	-
Lipid Droplet	-	-	-	-	-	-	-	-	-
Lysosome/Vacuole	-	-	-	-	-	-	-	-	-
Peroxisome	-	-	-	-	-	-	-	-	-
Vesicle	1.000	0.005	0.009	-	-	-	1.000	0.002	0.005
Primary Cilium	0.682	0.093	0.164	0.556	0.062	0.112	-	-	-
Secreted Proteins	0.903	0.761	0.826	0.877	0.730	0.797	0.604	0.338	0.433
Sperm	0.500	0.028	0.052	0.667	0.037	0.070	-	-	-
Micro Avg	0.690	0.386	0.495	0.693	0.382	0.492	0.598	0.236	0.338
Macro Avg	0.618	0.215	0.263	0.482	0.206	0.249	0.362	0.106	0.135
Subcellular	FoldSeek			Graph Transformer			Graph Mamba		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.616	0.398	0.484	0.664	0.543	0.597	0.562	0.556	0.559
Nuclear Membrane	-	-	-	-	-	-	0.061	0.026	0.037
Nucleoli	-	-	-	0.554	0.124	0.203	0.433	0.104	0.168
Nucleoplasm	0.591	0.341	0.433	0.642	0.483	0.552	0.526	0.481	0.502
Cytoplasm	0.581	0.102	0.174	0.552	0.305	0.393	0.476	0.373	0.418
Cytosol	0.500	0.001	0.003	0.457	0.170	0.248	0.421	0.431	0.426
Cytoskeleton	0.480	0.038	0.070	0.538	0.022	0.042	0.249	0.296	0.270
Centrosome	-	-	-	-	-	-	0.128	0.313	0.181
Mitochondria	-	-	-	0.688	0.363	0.475	0.407	0.294	0.341
Endoplasmic Reticulum	-	-	-	0.552	0.111	0.184	0.529	0.031	0.059
Golgi Apparatus	-	-	-	0.857	0.021	0.041	0.182	0.188	0.185
Cell Membrane	0.626	0.237	0.343	0.718	0.442	0.547	0.417	0.766	0.540
Endosome	-	-	-	-	-	-	0.125	0.083	0.100
Lipid Droplet	-	-	-	-	-	-	-	-	-
Lysosome/Vacuole	-	-	-	-	-	-	-	-	-
Peroxisome	-	-	-	-	-	-	-	-	-
Vesicle	-	-	-	0.526	0.023	0.044	0.306	0.086	0.135
Primary Cilium	-	-	-	0.500	0.006	0.012	0.205	0.056	0.088
Secreted Proteins	0.600	0.225	0.328	0.767	0.652	0.705	0.426	0.802	0.557
Sperm	-	-	-	1.000	0.009	0.018	0.116	0.147	0.130
Micro Avg	0.605	0.156	0.248	0.637	0.302	0.410	0.414	0.408	0.411
Macro Avg	0.200	0.067	0.092	0.451	0.164	0.203	0.279	0.252	0.235

Subcellular	CDConv[t]			GearNet-Edge[t]		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.651	0.592	0.620	0.619	0.450	0.521
Nuclear Membrane	-	-	-	-	-	-
Nucleoli	0.583	0.084	0.147	0.531	0.068	0.121
Nucleoplasm	0.633	0.541	0.583	0.613	0.444	0.515
Cytoplasm	0.580	0.414	0.483	0.498	0.491	0.495
Cytosol	0.489	0.277	0.353	0.417	0.358	0.385
Cytoskeleton	0.649	0.075	0.135	0.296	0.186	0.228
Centrosome	-	-	-	0.228	0.088	0.127
Mitochondria	0.707	0.359	0.476	0.470	0.240	0.318
Endoplasmic Reticulum	0.441	0.218	0.292	0.475	0.197	0.279
Golgi Apparatus	0.733	0.038	0.073	0.211	0.014	0.026
Cell Membrane	0.721	0.461	0.562	0.708	0.457	0.556
Endosome	-	-	-	0.364	0.037	0.067
Lipid Droplet	-	-	-	-	-	-
Lysosome/Vacuole	-	-	-	0.429	0.040	0.073
Peroxisome	-	-	-	-	-	-
Vesicle	0.667	0.014	0.027	0.270	0.039	0.068
Primary Cilium	-	-	-	0.467	0.087	0.147
Secreted Proteins	0.795	0.741	0.767	0.722	0.655	0.687
Sperm	-	-	-	0.714	0.046	0.086
Micro Avg	0.632	0.352	0.452	0.546	0.337	0.417
Macro Avg	0.382	0.191	0.226	0.402	0.195	0.235

[f]We finetune the pre-trained protein language model. [t]The original MLP is replaced by Transformer layers. [0]The parameters of ESM-C is initialized randomly. "/" indicates that DeepLoc 2.1 does not support prediction for that location, and therefore, average metrics are not considered in this case. "–" indicates that no prediction is made for that location.

Table 10: Detailed performance of selected baselines with reweighting scheme.

Subcellular	ESM-C 600M			CDConv[t]			GearNet-Edge[t]		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.698	0.575	0.630	0.481	0.892	0.625	0.484	0.856	0.618
Nuclear Membrane	-	-	-	0.033	0.566	0.062	0.046	0.079	0.058
Nucleoli	-	-	-	0.105	0.916	0.188	0.153	0.418	0.224
Nucleoplasm	0.679	0.500	0.576	0.469	0.859	0.607	0.436	0.841	0.574
Cytoplasm	0.568	0.446	0.500	0.450	0.823	0.582	0.441	0.711	0.544
Cytosol	0.513	0.076	0.133	0.353	0.829	0.495	0.366	0.714	0.484
Cytoskeleton	0.778	0.044	0.083	0.184	0.698	0.292	0.218	0.450	0.294
Centrosome	-	-	-	0.089	0.776	0.160	0.134	0.252	0.175
Mitochondria	0.846	0.336	0.481	0.191	0.672	0.297	0.247	0.427	0.313
Endoplasmic Reticulum	-	-	-	0.195	0.737	0.308	0.276	0.460	0.345
Golgi Apparatus	-	-	-	0.152	0.648	0.246	0.177	0.366	0.238
Cell Membrane	0.723	0.465	0.566	0.462	0.709	0.560	0.398	0.820	0.536
Endosome	-	-	-	0.067	0.407	0.114	0.177	0.130	0.150
Lipid Droplet	1.000	0.133	0.235	0.014	0.067	0.023	0.333	0.067	0.111
Lysosome/Vacuole	-	-	-	0.117	0.347	0.175	0.116	0.107	0.111
Peroxisome	1.000	0.105	0.190	0.040	0.421	0.072	0.111	0.105	0.108
Vesicle	-	-	-	0.198	0.532	0.288	0.206	0.445	0.281
Primary Cilium	0.667	0.012	0.024	0.096	0.640	0.167	0.123	0.311	0.176
Secreted Proteins	0.833	0.730	0.778	0.413	0.891	0.564	0.509	0.775	0.614
Sperm	-	-	-	0.066	0.679	0.120	0.109	0.147	0.125
Micro Avg	0.679	0.313	0.429	0.253	0.772	0.381	0.348	0.650	0.453
Macro Avg	0.415	0.171	0.210	0.209	0.655	0.197	0.253	0.424	0.304

[t]The original MLP is replaced by Transformer layers. "–" indicates that no prediction is made for that location.

Table 11: Detailed performance of selected baselines with single-label classification strategy.

Subcellular	ESM-C 600M			CDConv[t]			GearNet-Edge[t]		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nuclear Membrane	-	-	-	0.027	0.711	0.052	0.026	0.118	0.042
Nucleoli	0.251	0.285	0.267	0.082	0.992	0.151	0.151	0.470	0.228
Centrosome	0.124	0.361	0.184	0.051	0.333	0.089	0.099	0.531	0.167
Golgi Apparatus	0.293	0.268	0.280	0.080	0.199	0.114	0.161	0.303	0.210
Endosome	0.111	0.333	0.167	0.029	0.176	0.049	0.082	0.278	0.126
Lipid Droplet	0.011	0.200	0.021	-	-	-	0.032	0.133	0.051
Lysosome/Vacuole	0.075	0.253	0.115	-	-	-	0.097	0.493	0.162
Peroxisome	0.029	0.526	0.054	-	-	-	0.013	0.158	0.023
Vesicle	0.270	0.039	0.068	0.141	0.625	0.230	0.207	0.380	0.268
Primary Cilium	0.175	0.460	0.253	0.055	0.379	0.097	0.104	0.472	0.171
Sperm	0.121	0.229	0.159	0.045	0.138	0.068	0.077	0.239	0.117

[t]The original MLP is replaced by Transformer layers. "–" indicates that no prediction is made for that location.

Table 12: Detailed performance comparison of CDConv and GearNet-Edge under random sampling of Cα coordinates.

Subcellular	CDConv[t] (ablation)			CDConv[t]			GearNet-Edge[t] (ablation)			GearNet-Edge[t]		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.595	0.512	0.550	0.651	0.592	0.620	0.515	0.459	0.485	0.619	0.450	0.521
Nuclear Membrane	-	-	-	-	-	-	-	-	-	-	-	-
Nucleoli	0.417	0.020	0.038	0.583	0.084	0.147	0.268	0.076	0.119	0.531	0.068	0.121
Nucleoplasm	0.578	0.419	0.486	0.633	0.541	0.583	0.479	0.428	0.452	0.613	0.444	0.515
Cytoplasm	0.535	0.214	0.306	0.580	0.414	0.483	0.432	0.414	0.422	0.498	0.491	0.495
Cytosol	0.478	0.069	0.120	0.489	0.277	0.353	0.394	0.279	0.327	0.417	0.358	0.385
Cytoskeleton	-	-	-	0.649	0.075	0.135	0.252	0.119	0.162	0.296	0.186	0.228
Centrosome	-	-	-	-	-	-	0.184	0.061	0.092	0.228	0.088	0.127
Mitochondria	0.537	0.084	0.145	0.707	0.359	0.476	0.283	0.065	0.106	0.470	0.240	0.318
Endoplasmic Reticulum	0.625	0.017	0.034	0.441	0.218	0.292	0.321	0.062	0.104	0.475	0.197	0.279
Golgi Apparatus	-	-	-	0.733	0.038	0.073	0.132	0.017	0.031	0.211	0.014	0.026
Cell Membrane	0.621	0.425	0.505	0.721	0.461	0.562	0.595	0.413	0.487	0.708	0.457	0.556
Endosome	-	-	-	-	-	-	-	-	-	0.364	0.037	0.067
Lipid Droplet	-	-	-	-	-	-	-	-	-	-	-	-
Lysosome/Vacuole	-	-	-	-	-	-	-	-	-	0.429	0.040	0.073
Peroxisome	-	-	-	-	-	-	-	-	-	-	-	-
Vesicle	-	-	-	0.667	0.014	0.027	0.185	0.077	0.109	0.270	0.039	0.068
Primary Cilium	-	-	-	-	-	-	0.368	0.043	0.078	0.467	0.087	0.147
Secreted Proteins	0.703	0.218	0.333	0.795	0.741	0.767	0.515	0.232	0.320	0.722	0.655	0.687
Sperm	-	-	-	-	-	-	0.125	0.009	0.017	0.714	0.046	0.086
Micro Avg	0.586	0.229	0.329	0.632	0.352	0.452	0.450	0.283	0.348	0.546	0.337	0.417
Macro Avg	0.254	0.099	0.126	0.382	0.191	0.226	0.252	0.138	0.166	0.402	0.195	0.235

[t]The original MLP is replaced by Transformer layers. "–" indicates that no prediction is made for that location.

we conduct an ablation study on two representative structure-based baselines to quantify the positive impact of 3D information incorporated.

Specifically, to preserve the integrity of the model input, we performed preprocessing on the protein structural data. For each protein, we obtained the boundary values of its 3D coordinates and uniformly sampled the Cα coordinates at random within these boundaries to generate new protein structures. The 1D sequence data were kept unchanged, while the randomly sampled structures were used as the 3D structural input. Using the same hyperparameter settings as in the main experiments, we conducted an ablation study, with the detailed results shown in Table 12. We observed a significant performance drop in this setting, which further demonstrates the decisive role of accurate 3D structural input in enabling correct model predictions.

F Explanation and Illustrative Examples of Evidence-level Annotations

The evidence-level annotations design was originally intended to allow researchers to flexibly select annotations based on their specific use cases. For instance, when the goal is to identify subcellular localization signals with high precision, selecting annotations with high confidence (*i.e.,* choosing the experimentally validated annotations only) is more appropriate. Conversely, for large-scale protein localization prediction, using lower-confidence but more abundant annotations (*i.e.,* choosing both the non-experimentally validated and non-experimentally validated annotations) may lead to better model performance.

In our main experiments, all non-experimentally validated annotations were treated as positive samples to enhance the models' performance in high-throughput prediction settings. Here we present two illustrative examples of the flexible usages of evidence-level annotations: 1) **weighting labels** (*i.e.,* treating non-experimentally validated annotations as positive samples, but assigning a weight of 0.7 to them relative to experimental ones, which reduces the weight of non-experimental data in influencing the model) and 2) **filtering labels** (*i.e.,* treating non-experimentally validated annotations as negative samples, which restricts models learning to experimental data with high reliability). The results are compared with the original one in our paper in Table 13.

As shown in the table, models that treat non-experimentally validated annotations as positive samples generally achieve the best overall performance. This may be because many of the non-experimentally validated annotations in the UniProt database are derived from biological papers; thus, they still hold relatively high credibility. This also demonstrates that including non-experimentally validated annotations in the dataset can be beneficial for helping the models capture meaningful localization signals and patterns. In contrast, down-weighting these annotations or even treating them as negative samples tends to degrade the models' overall performance.

Table 13: Detailed performance of two illustrative examples of evidence-level annotations: weighting labels and filtering labels.

Subcellular	ESM-C 600M			ESM-C 600M (weighting)			ESM-C 600M (filtering)		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.694	0.609	0.649	0.717	0.585	0.645	0.703	0.575	0.633
Nuclear Membrane	-	-	-	-	-	-	-	-	-
Nucleoli	0.800	0.048	0.091	0.786	0.088	0.159	0.867	0.055	0.104
Nucleoplasm	0.679	0.573	0.621	0.692	0.558	0.618	0.686	0.551	0.611
Cytoplasm	0.611	0.477	0.536	0.619	0.453	0.523	0.572	0.400	0.470
Cytosol	0.541	0.307	0.392	0.534	0.256	0.346	0.590	0.239	0.340
Cytoskeleton	0.681	0.154	0.251	0.649	0.116	0.197	0.381	0.034	0.062
Centrosome	1.000	0.007	0.014	1.000	0.007	0.014	-	-	-
Mitochondria	0.865	0.416	0.562	0.907	0.374	0.530	0.798	0.373	0.508
Endoplasmic Reticulum	0.687	0.235	0.351	0.726	0.156	0.256	0.500	0.039	0.072
Golgi Apparatus	0.938	0.052	0.099	1.000	0.010	0.021	-	-	-
Cell Membrane	0.777	0.531	0.631	0.757	0.570	0.650	0.661	0.254	0.367
Endosome	1.000	0.009	0.018	-	-	-	-	-	-
Lipid Droplet	-	-	-	-	-	-	-	-	-
Lysosome/Vacuole	-	-	-	-	-	-	-	-	-
Peroxisome	-	-	-	-	-	-	-	-	-
Vesicle	1.000	0.005	0.009	1.000	0.005	0.009	-	-	-
Primary Cilium	0.682	0.093	0.164	0.538	0.043	0.080	1.000	0.008	0.016
Secreted Proteins	0.903	0.761	0.826	0.920	0.669	0.775	-	-	-
Sperm	0.500	0.028	0.052	0.800	0.037	0.070	0.500	0.010	0.019
Micro Avg	0.690	0.386	0.495	0.700	0.366	0.481	0.657	0.306	0.418
Macro Avg	0.618	0.215	0.263	0.582	0.196	0.245	0.363	0.127	0.160
Subcellular	CDConv[t]			CDConv[t] (weighting)			CDConv[t] (filtering)		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.651	0.592	0.620	0.674	0.539	0.599	0.644	0.530	0.581
Nuclear Membrane	-	-	-	-	-	-	-	-	-
Nucleoli	0.583	0.084	0.147	0.556	0.020	0.039	0.484	0.640	0.112
Nucleoplasm	0.633	0.541	0.583	0.657	0.497	0.566	0.652	0.424	0.514
Cytoplasm	0.580	0.414	0.483	0.557	0.469	0.509	0.528	0.445	0.483
Cytosol	0.489	0.277	0.353	0.470	0.293	0.361	0.468	0.337	0.392
Cytoskeleton	0.649	0.075	0.135	0.714	0.063	0.116	0.400	0.008	0.016
Centrosome	-	-	-	-	-	-	-	-	-
Mitochondria	0.707	0.359	0.476	0.762	0.355	0.484	0.588	0.363	0.449
Endoplasmic Reticulum	0.441	0.218	0.292	0.561	0.159	0.248	0.278	0.024	0.045
Golgi Apparatus	0.733	0.038	0.073	0.615	0.028	0.053	-	-	-
Cell Membrane	0.721	0.461	0.562	0.723	0.447	0.553	0.562	0.194	0.288
Endosome	-	-	-	-	-	-	-	-	-
Lipid Droplet	-	-	-	-	-	-	-	-	-
Lysosome/Vacuole	-	-	-	-	-	-	-	-	-
Peroxisome	-	-	-	-	-	-	-	-	-
Vesicle	0.667	0.014	0.027	-	-	-	-	-	-
Primary Cilium	-	-	-	0.333	0.006	0.012	-	-	-
Secreted Proteins	0.795	0.741	0.767	0.857	0.573	0.687	0.400	0.044	0.079
Sperm	-	-	-	-	-	-	-	-	-
Micro Avg	0.632	0.352	0.452	0.637	0.333	0.438	0.577	0.291	0.386
Macro Avg	0.382	0.191	0.226	0.374	0.173	0.211	0.250	0.122	0.148
Subcellular	GearNet-Edge[t]			GearNet-Edge[t] (weighting)			GearNet-Edge[t] (filtering)		
Locations	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Nucleus	0.619	0.450	0.521	0.622	0.337	0.437	0.607	0.478	0.535
Nuclear Membrane	-	-	-	-	-	-	-	-	-
Nucleoli	0.531	0.068	0.121	0.421	0.032	0.060	0.333	0.064	0.107
Nucleoplasm	0.613	0.444	0.515	0.618	0.326	0.427	0.576	0.453	0.507
Cytoplasm	0.498	0.491	0.495	0.473	0.466	0.469	0.452	0.479	0.465
Cytosol	0.417	0.358	0.385	0.424	0.339	0.377	0.394	0.403	0.398
Cytoskeleton	0.296	0.186	0.228	0.342	0.167	0.224	0.284	0.105	0.153
Centrosome	0.228	0.088	0.127	0.213	0.068	0.103	0.200	0.054	0.085
Mitochondria	0.470	0.240	0.318	0.667	0.176	0.278	0.508	0.142	0.221
Endoplasmic Reticulum	0.475	0.197	0.279	0.435	0.128	0.198	0.268	0.073	0.115
Golgi Apparatus	0.211	0.014	0.026	0.211	0.014	0.026	0.250	0.004	0.008
Cell Membrane	0.708	0.457	0.556	0.629	0.392	0.483	0.463	0.170	0.248
Endosome	0.364	0.037	0.067	0.333	0.028	0.051	0.500	0.029	0.056
Lipid Droplet	-	-	-	-	-	-	-	-	-
Lysosome/Vacuole	0.429	0.040	0.073	0.125	0.013	0.024	-	-	-
Peroxisome	-	-	-	-	-	-	-	-	-
Vesicle	0.270	0.039	0.068	0.268	0.093	0.138	0.280	0.055	0.092
Primary Cilium	0.467	0.087	0.147	0.524	0.068	0.121	0.364	0.031	0.058
Secreted Proteins	0.722	0.655	0.687	0.826	0.519	0.637	0.273	0.066	0.106
Sperm	0.714	0.046	0.086	0.667	0.037	0.070	0.500	0.020	0.038
Micro Avg	0.546	0.337	0.417	0.529	0.282	0.368	0.485	0.300	0.371
Macro Avg	0.402	0.195	0.235	0.390	0.160	0.206	0.313	0.131	0.160

[t]The original MLP is replaced by Transformer layers. "–" indicates that no prediction is made for that location.

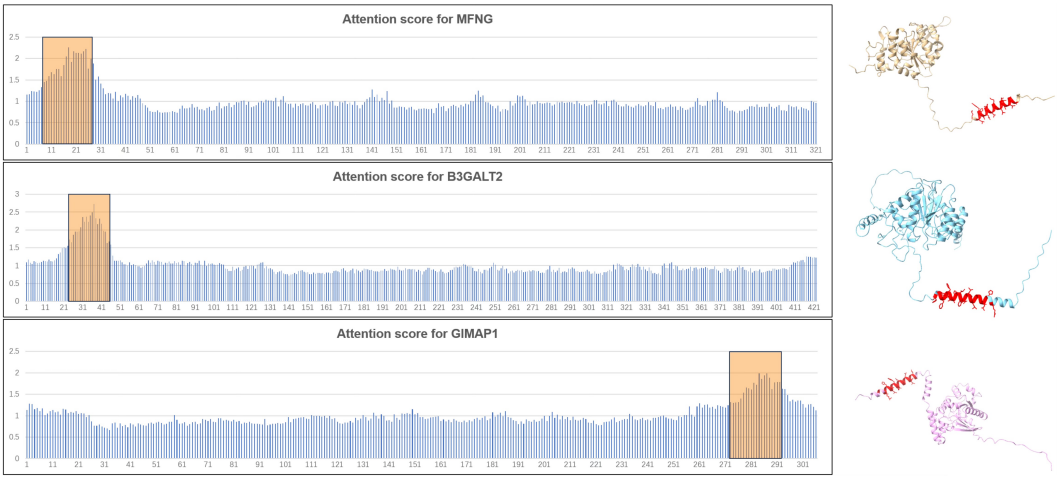

Figure 3: Visualization of full attention scores and structures of proteins MFNG, B3GALT2, and GIMAP1, where the residues of known pattern α-helix are highlighted.

However, we found that as the annotation confidence increased (*i.e.,* from treating non-experimentally validated annotations as positive samples, to lowering their weights, and finally to treating them as negative samples), the precision of the models generally improved, especially for those subcellular locations with more positive samples. Actually, precision and recall often represent a trade-off in modeling strategies. That is, adopting a more conservative prediction strategy typically increases precision but reduces the number of correctly recalled samples, and vice versa. Therefore, selecting high-confidence evidence levels can be seen as a method of enforcing a more conservative prediction approach, helping to reduce the likelihood of false-positive predictions. This highlights the novelty of evidence-level annotations: using experimentally validated data helps ensure models' high precision and confidence.

G Interpretability with Attention Score

In Transformer architectures, the attention mechanism allows each token to compute a weighted representation of all other tokens in the sequence. Specifically, for a given token, a set of attention weights is derived via scaled dot-product operations between its query vector and the key vectors of all tokens, followed by a softmax normalization. These attention weights reflect how much information the token attends to from each of its peers. To assess the relative importance of each token within the sequence, we aggregated the attention it receives from all other tokens, *i.e.,* summing over the attention scores directed toward that token across the entire sequence. This provides a global measure of how influential a token is in shaping the contextual representations learned by the model. We interpret this aggregated attention as a proxy for biological interpretability, where highly attended residues may correspond to structurally or functionally important positions within the protein.

In Section 4.3.3 of the main text, we introduce a CDConv model for predicting Golgi apparatus localization. By analyzing the attention score within the model's Transformer architecture, we identify a localization pattern associated with an α-helix, which is consistent with existing biological findings. Here, we visualize the full attention score of the three example proteins discussed in the main text (*i.e.,* MFNG, B3GALT2, and GIMAP1), as shown in Figure 3. The residues of known localization patterns α-helix are highlighted in orange for clear comparison. Notably, the 20 residues with the highest attention scores exhibit a 90% overlap with the ground truth, further highlighting the CDConv model's precision in identifying localization patterns.

H Availability of Dataset and Code

The complete dataset, including localization labels, extracted protein structures, *etc.* can be accessed at https://huggingface.co/datasets/getbetterhyccc/CAPSUL. Our implemen-

tation is publicly available at https://anonymous.4open.science/r/CAPSUL-37E2. For some baseline models, we adopt publicly released implementations, including Graph Transformer at https://github.com/pyg-team/pytorch_geometric/tree/master and Graph Mamba at https://github.com/alxndrTL/mamba.py.

I THE USE OF LARGE LANGUAGE MODELS

In this study, Large Language Models (LLMs) were employed solely for linguistic refinement, such as polishing the clarity, grammar, and fluency of the manuscript. Importantly, all conceptual advances, methodological innovations, experimental designs, and primary contributions presented in this work were independently conceived, developed, and validated by the authors. The role of LLMs was thus limited to improving readability and ensuring the precision of academic writing, without influencing the scientific content or originality of the research.