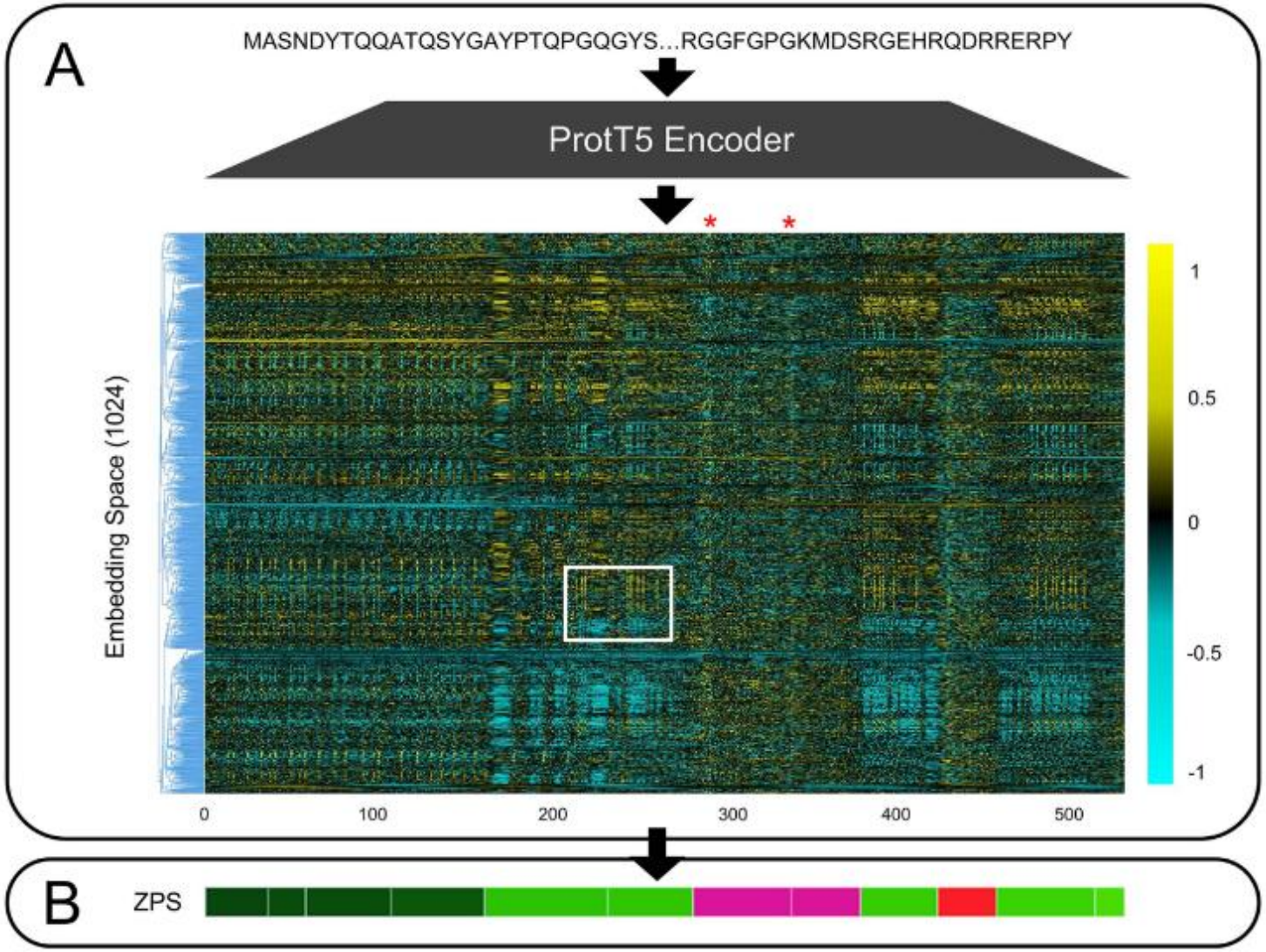
Zero-shot segmentation using embeddings from a protein language model identifies functional regions in the human proteome

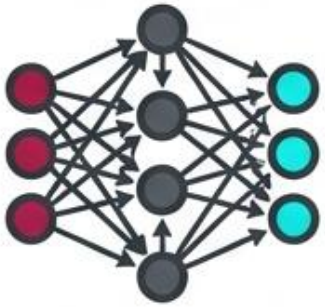
Ami G. Sangster, Cameron Dufault, Haoning Qu, Denise Le, Julie D. Forman-Kay, Alan M. Moses

PLOS Computational Biology, 2025



Executive Summary: The Unsupervised Discovery Arc

The Innovation



- Uses **ProtT5-XL-UniRef50** embeddings (1024-dim) in JetBrains Mono.
- **Zero-shot**: No fine-tuning or training on segmentation tasks.
- **Method**: **Change Point Analysis** detects shifts in embedding signals.

The Validation

IoU and Boundary Evaluation

- **IDRs**: **Outperforms** unsupervised baselines (fLPS2, Chi-Score).
- **Domains**: **Competitive** with supervised methods (**Pfam**) on boundaries.

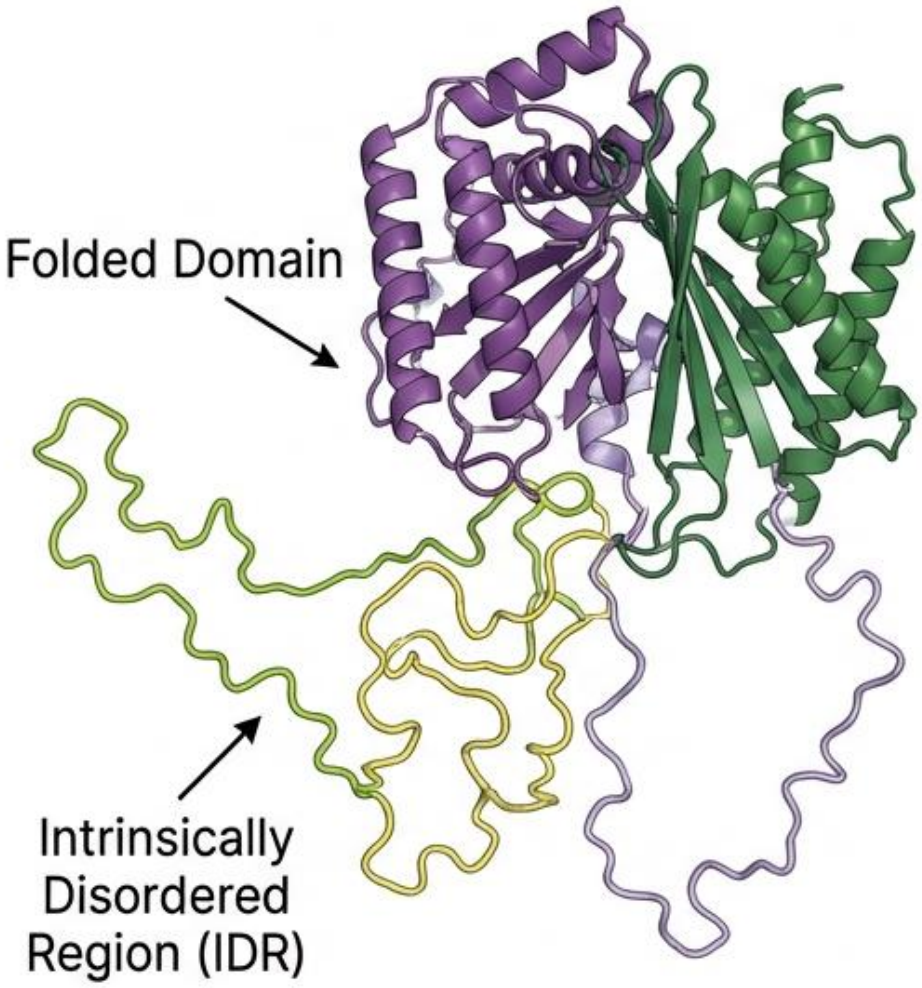
1-NN Classification

- **Categorization**: Distinguishes >200 **UniProt** annotation types via embedding similarity.

The Discovery

- **Mitochondrial Signals**: Identified 51 unannotated targeting signals.
- **"Stealth" Domains**: Discovered **prion-like** (SYGQ-rich) domains missed by BLAST.
- **Resolution**: Solves '**unknown length**' annotations in the human proteome.

The Segmentation Bottleneck: When Sequence Conservation Fails



- **The Challenge:** Conventional bioinformatics tools (e.g., Pfam, Prosite) rely on evolutionary sequence alignment.
- **The Gap:** This approach works well for Folded Domains, which have stable structures and conserved sequences. However it **fails** for intrinsically disordered Regions (IDRs), which **lack stable structure** yet are present in **>60% of human proteins**
- **Current Limits:**
  - Supervised Learning requires labeled data, limiting discovery of new regions.
  - Existing unsupervised methods (e.g., fLPS2) focus only on compositional bias (e.g., 'rich in Glycine'), missing the broader functional context.

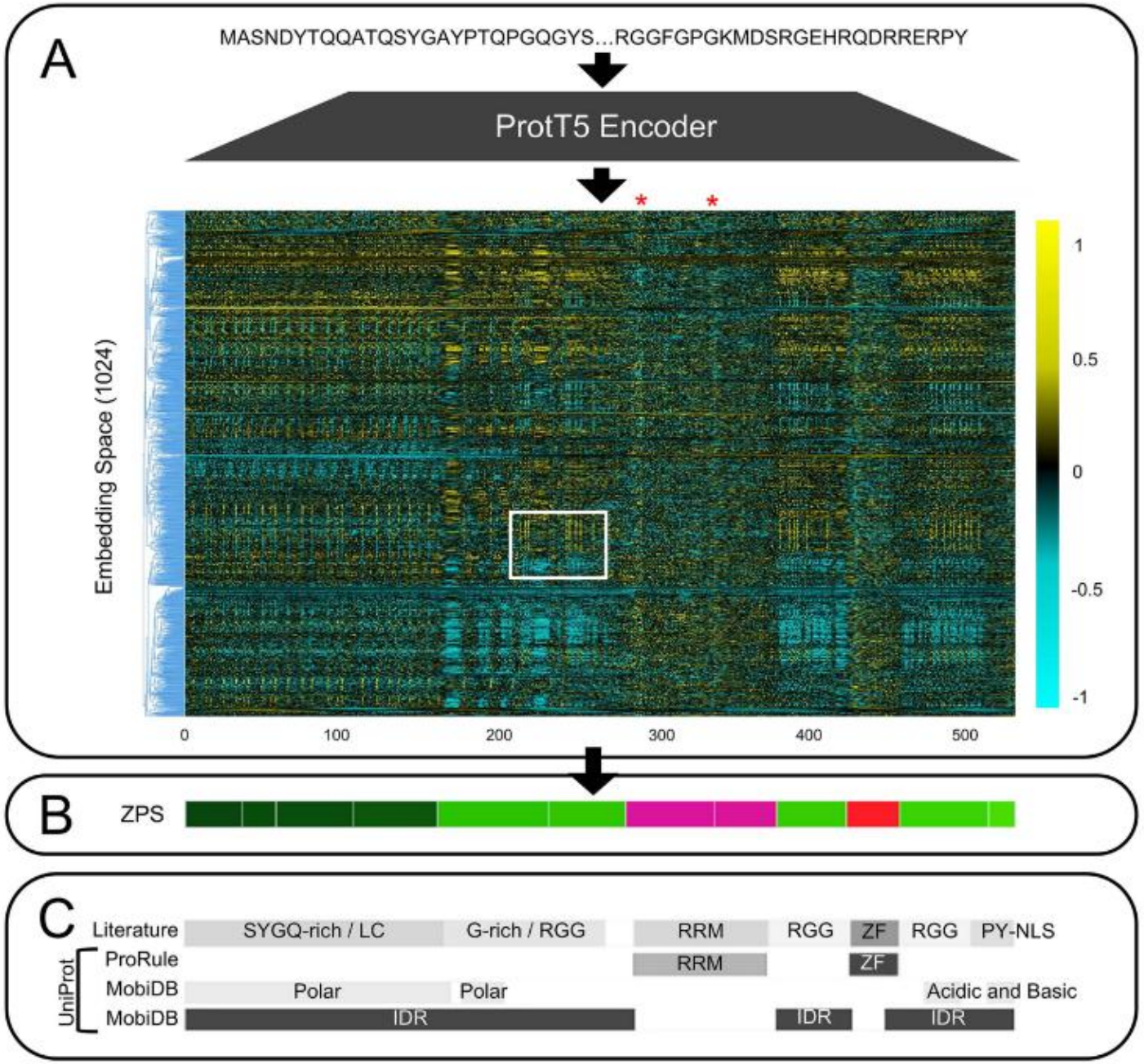
Pipeline Overview

ZPS (Zero-Shot Protein Segmentation)

Segment embeddings: average pooling the segmented embedding to 1x1024

Correction for over-segmentation: The cosine similarity between all segments in a protein are measured; if the segments with the highest cosine similarity are adjacent to each other in the protein sequence, then they are merged. (only if there are 6 or more segments)

Segment embedding to RGB colors: They use a UMAP to 3 dimensions. Then they scale these by mean-centering + sigmoid + multiplication by 255. Each dimension is represented as a different base color (RGB).



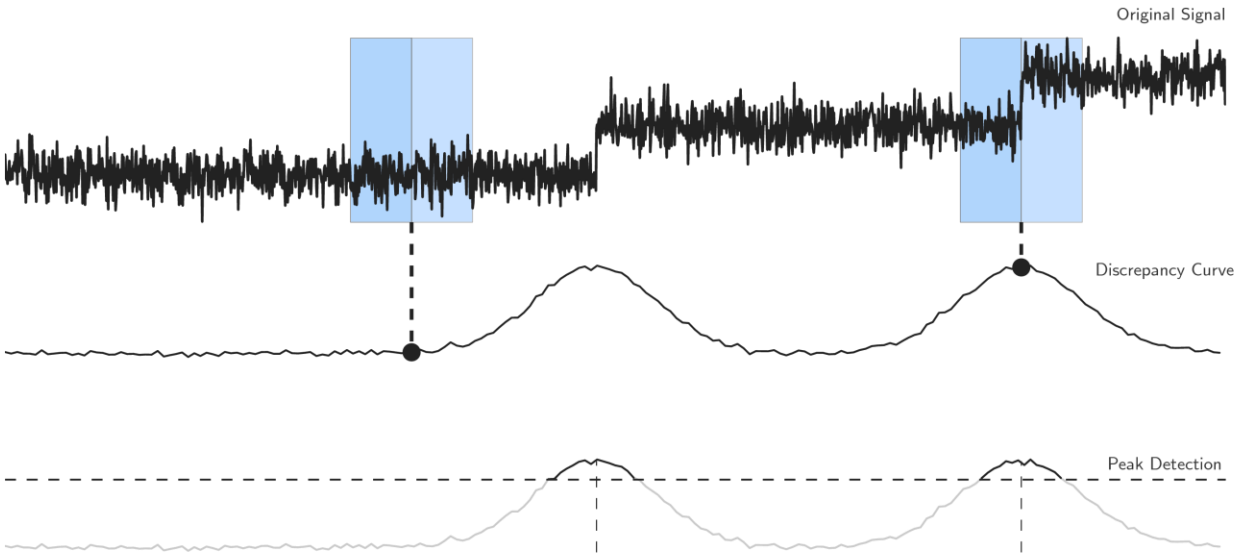
Protein: human protein FUS (UniProt ID: P35637)

Segmentation Method

Change Point Analysis

The statistical properties of the signals within each window are compared with a discrepancy measure. For a given cost function $c(\cdot)$, a discrepancy measure is derived $d(\cdot, \cdot)$ as follows:

$$d(y_{u..v}, y_{v..w}) = c(y_{u..w}) - c(y_{u..v}) - c(y_{v..w})$$



Source: ruptures Python package documentation (https://centre-borelli.github.io/ruptures-docs/user-guide/detection/window)

Hyperparameters:

Cost function: RBF Kernel
Window size: 30
Maximum of 3 change points per 100 residues

Cost function 12 ($c_{\mathbf{rbf}}$). *The cost function c_{rbf} is given by*

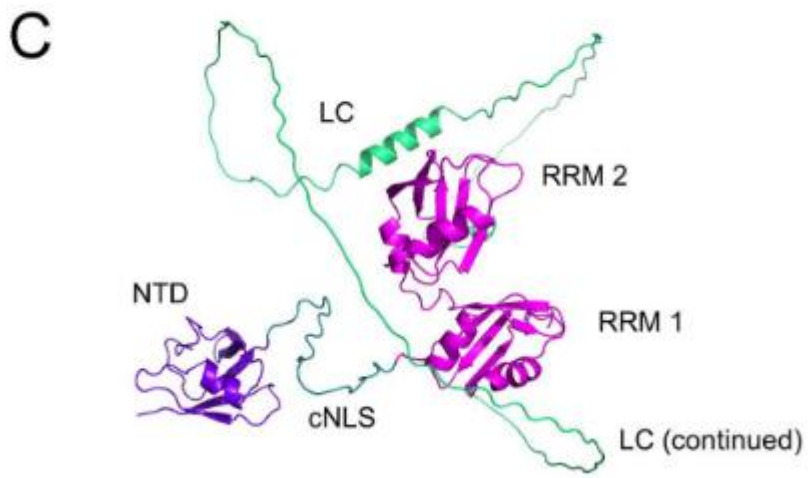
$$c_{rbf}(y_{a..b}) := (b-a) - \frac{1}{b-a} \sum_{s,t=a+1}^{b} \exp(-\gamma \|y_s - y_t\|^2)$$

where $\gamma > 0$ is the so-called bandwidth parameter.

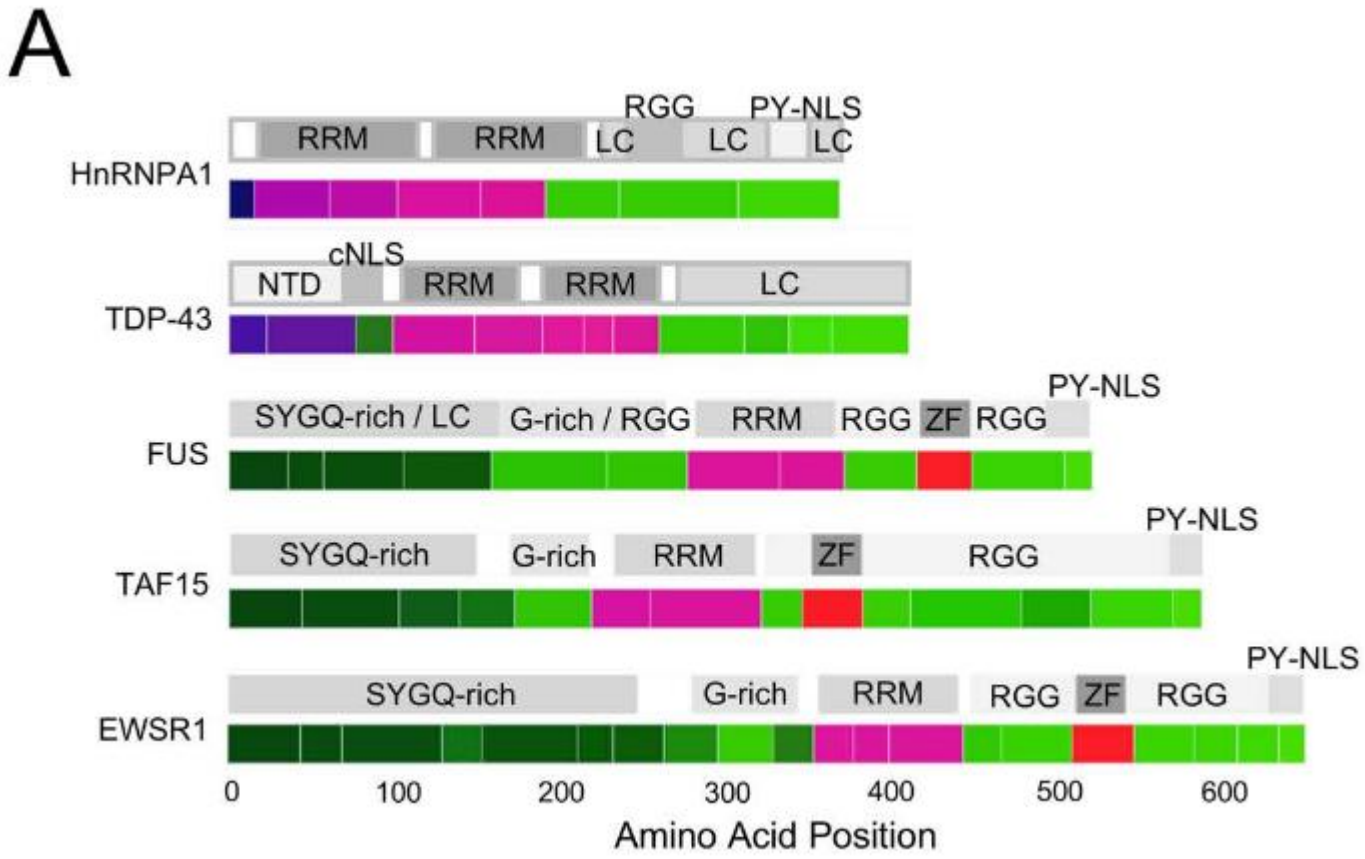
Case Study

Similar Segmentations for a Protein Family

FET family (FUS, EWSR1, and TAF15), TDP-43, and HnRNPA1 was used to visualize segment embedding. They found similar colors in same annotations across proteins.



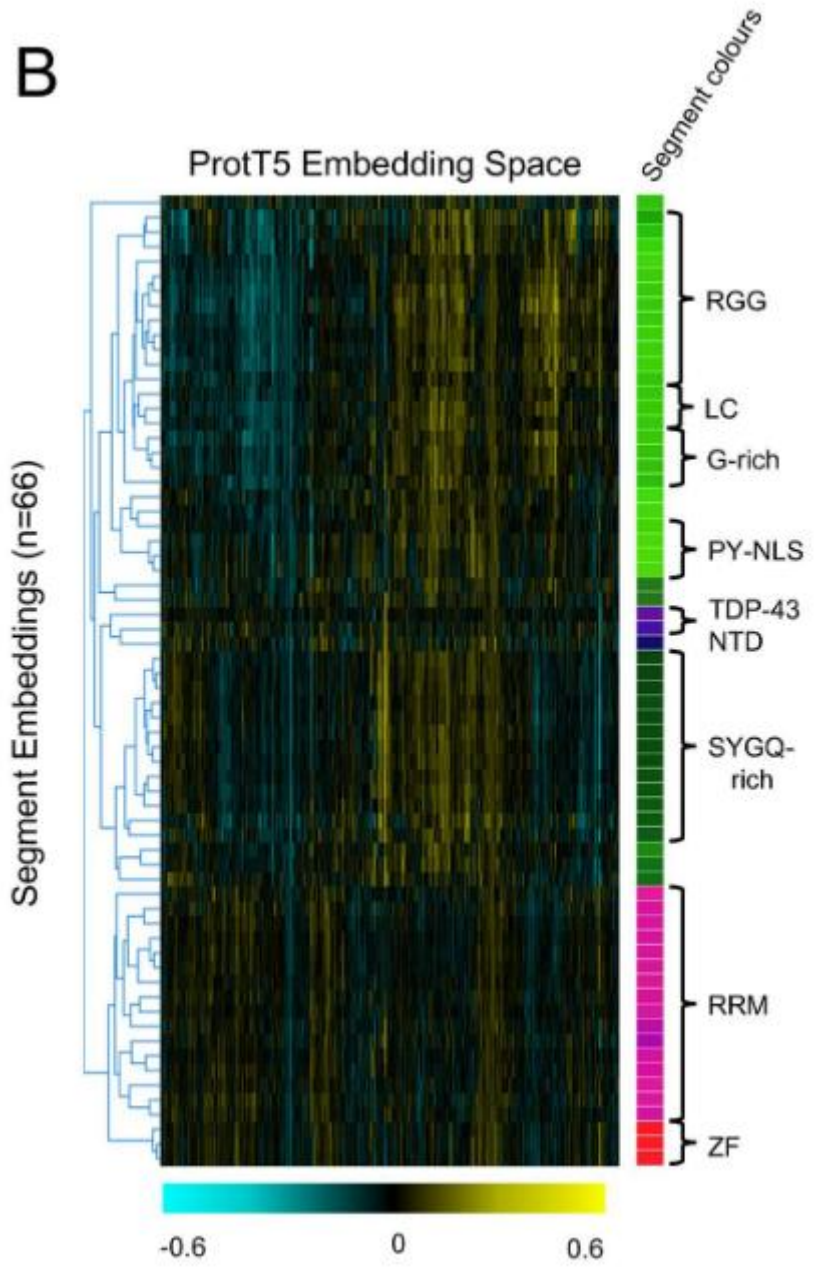
TDP-43



Case Study

Similar Segmentations for a Protein Family

When the high-dimensional segment embeddings were clustered hierarchically they also found similar color but different functional regions (RGG and PY-NLS) were clustered seperately.



Segmentation Evaluation

IoU Calculation

Table 1. Segmentation Evaluation for Human Proteome Segment Annotations from UniProt.

	All UniProt (N=92k)					MobiDB (N=23k)					Prosite (N=22k)				
	Pred. seg.	AloU (pred)	AloU (pos)	Prec.	Rec.	Pred. seg.	AloU (pred)	AloU (pos)	Prec.	Rec.	Pred. seg.	AloU (pred)	AloU (pos)	Prec.	Rec.
Supervised															
Pfam	59k	0.466	0.364	0.497	0.320	29k	0.033	0.047	0.018	0.023	40k	0.388	0.711	0.435	0.793
Prosite Scan	42k	**0.638**	0.363	**0.635**	0.289	24k	0.111	0.137	0.097	0.098	33k	**0.570**	**0.879**	**0.576**	**0.853**
Unsupervised															
fLPS2	109k	0.268	0.342	0.204	0.242	68k	0.152	0.382	0.112	0.326	61k	0.123	0.299	0.075	0.205
fLPS2 (low complexity)	48k	0.247	0.148	0.198	0.102	37k	**0.206**	0.328	**0.171**	0.271	28k	0.055	0.079	0.043	0.053
Chi-Score Analysis	554k	0.107	0.340	0.036	0.215	317k	0.051	0.429	0.026	0.352	310k	0.054	0.278	0.009	0.120
Chi-Score (filtered)	81k	0.336	0.375	0.299	0.265	51k	0.177	0.391	0.146	0.316	43k	0.161	0.332	0.129	0.253
Our Method															
ZPS	253k	0.231	**0.525**	0.172	**0.472**	152k	0.107	**0.580**	0.088	**0.569**	144k	0.123	0.534	0.077	0.497
ZPS (corrected)	164k	0.282	0.509	0.230	0.411	98k	0.129	0.527	0.111	0.466	93k	0.158	0.541	0.120	0.505

We report the number of predicted segments (Pred. seg.) and Intersection over Union (IoU) metrics such as Average IoU (AIoU) with respect to the predicted (pred) and positive (pos) annotations (see Methods) Precision (Prec.), and Recall (Rec.). ZPS (corrected) uses a simple correction for over-segmentation (see Methods). The best performance for each measure is shown in **bold.** N shows the number of annotated segments in the dataset. We use an IoU threshold of 0.5 to define precision and recall.

Boundary Evaluation

Closeness to the boundaries of highest IoU

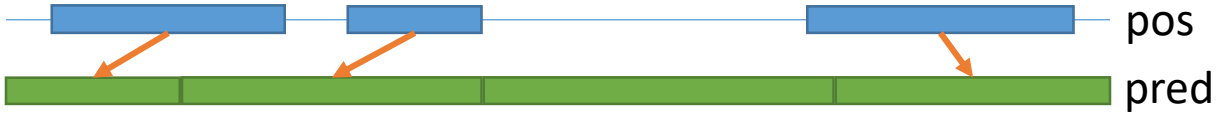


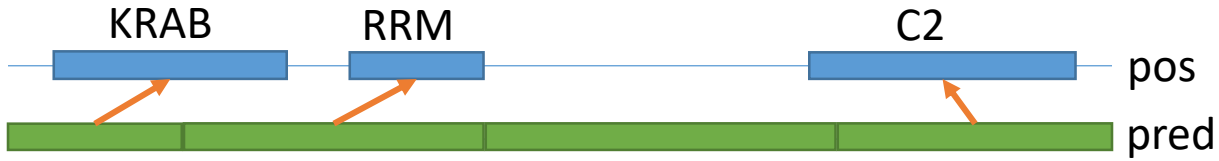
Table 2. Boundary Evaluation for Human Proteome Segment Annotations from UniProt.

	All Uniprot (N=92k)			MobiDB (N=23k)			ProRule (N=22k)		
	Predicted Segments	Missed Uniprot Segments	<10 amino acids	Predicted Segments	Missed MobiDB Segments	<10 amino acids	Predicted Segments	Missed ProRule Segments	<10 amino acids
Supervised									
Pfam	59k	31k	26.31%	29k	18k	2.40%	40k	1k	58.02%
Prosite Scan	42k	38k	31.23%	24k	15k	9.44%	33k	0.5k	**82.75%**
Unsupervised									
fLPS2	109k	7k	19.15%	68k	0.9k	27.72%	61k	2k	14.06%
fLPS2 (low complexity)	48k	52k	12.51%	37k	6k	28.42%	28k	16k	6.92%
Chi-Score Analysis	554k	11k	28.06%	317k	3k	37.14%	310k	4k	19.98%
Chi-Score (filtered)	81k	11k	23.40%	51k	3k	30.07%	43k	4k	15.84%
Our Method									
ZPS	253k	1k	**41.36%**	152k	0.3k	**48.51%**	144k	0.3k	38.62%
ZPS (corrected)	164k	1k	37.08%	98k	0.3k	41.32%	93k	0.3k	35.61%

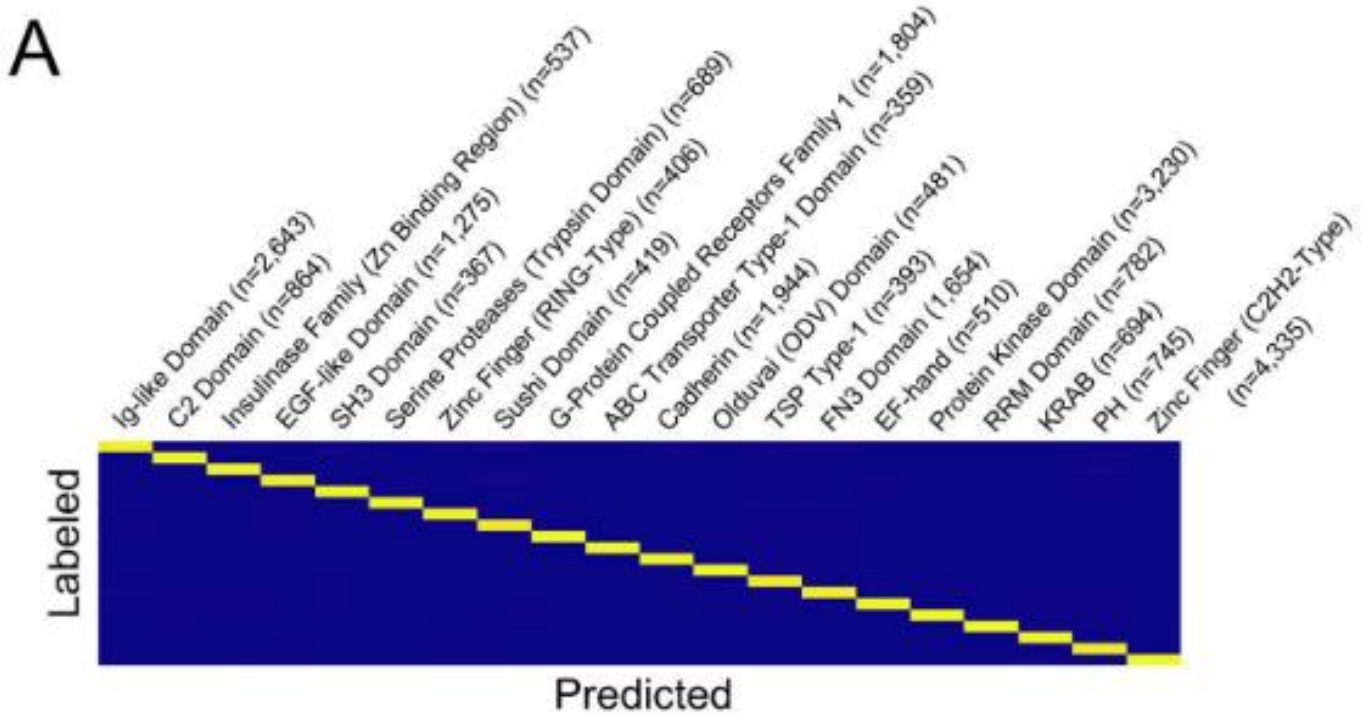
We report the number of predicted segments, the number of annotated segments without an overlapping predicted segment (Unpaired Segments), and the percentage of annotated boundaries with a predicted boundary less than 10 amino acids away. ZPS (corrected) uses a simple correction for over-segmentation (see Methods). The best performance for each measure is shown in **bold.** N shows the number of annotated segments in the dataset.

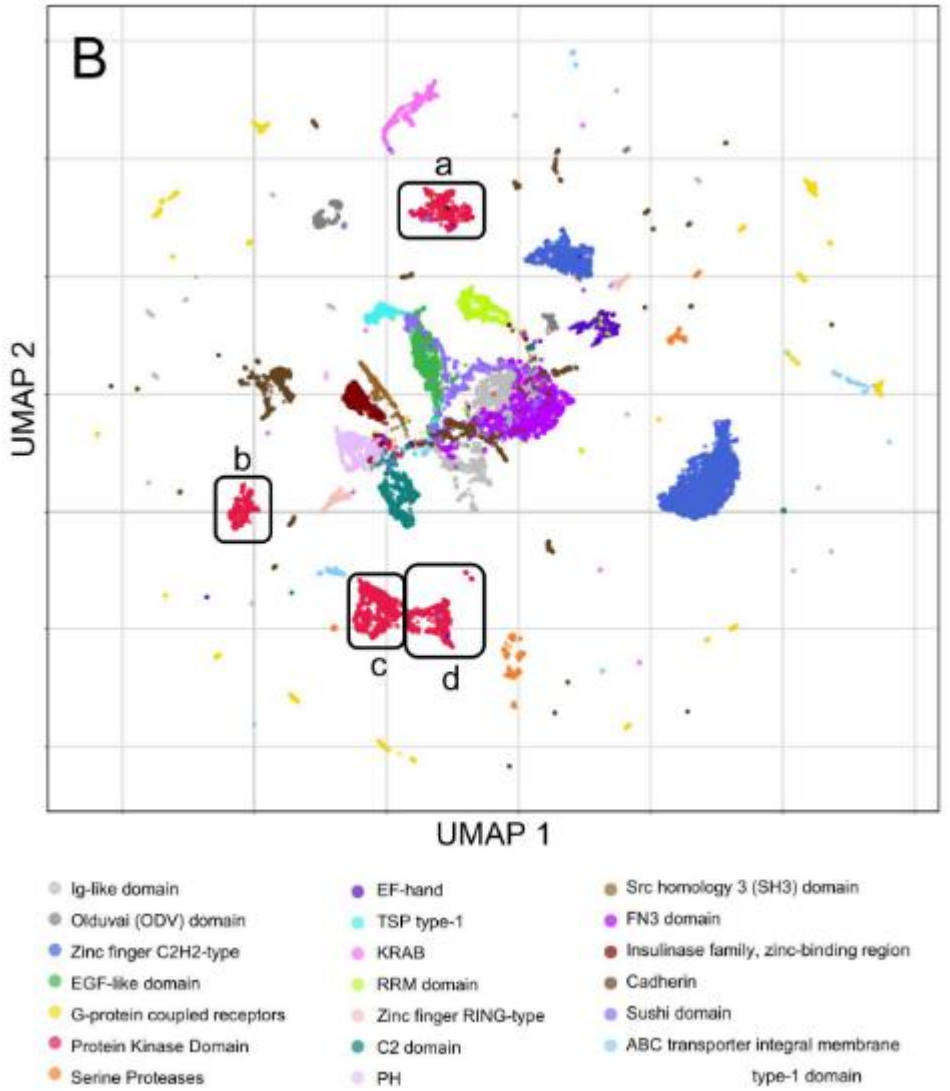
1-NN Folded Domain Classification

Top 20 ProRule Domains



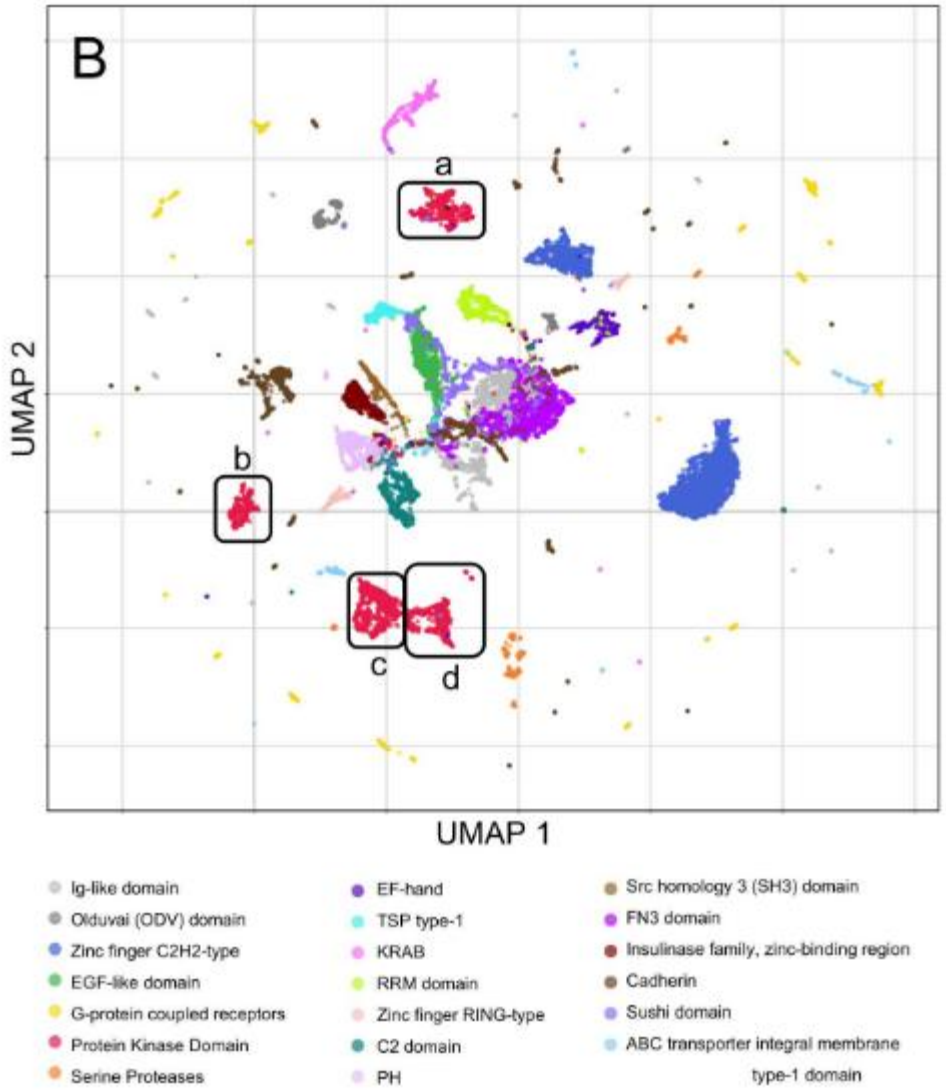
1-nn precisions ranging from 0.951 to 1 (+/- 0.004 to ~0) and an average precision of 0.986 + /- 0.002

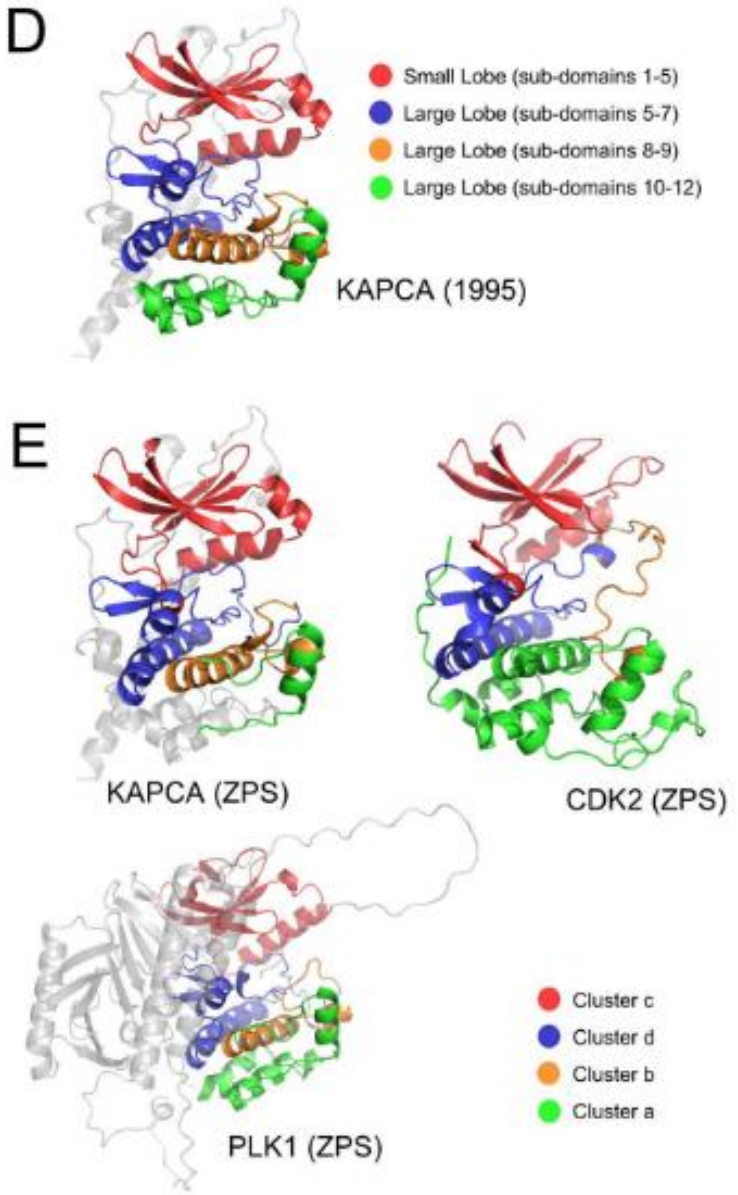




1-NN Folded Domain Classification

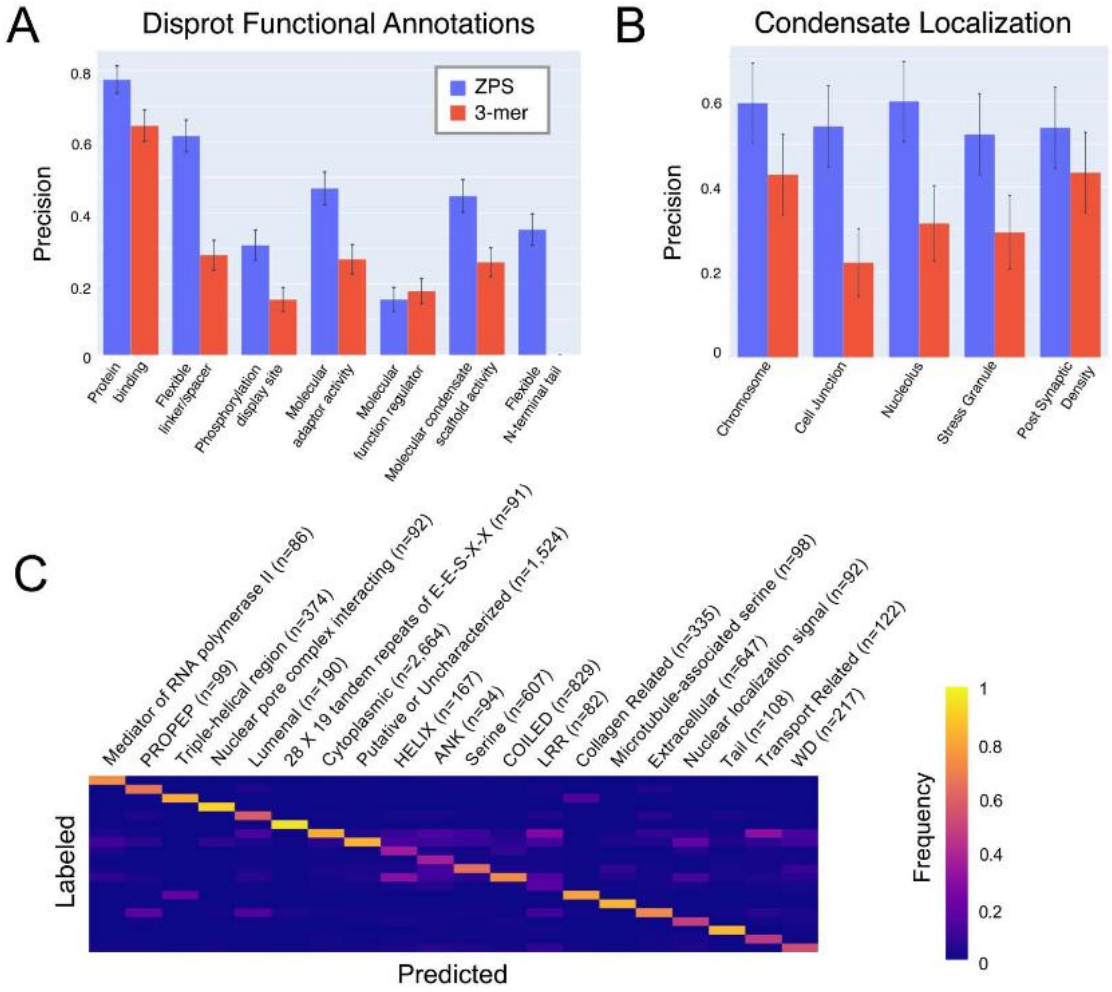
SubDomain Finding

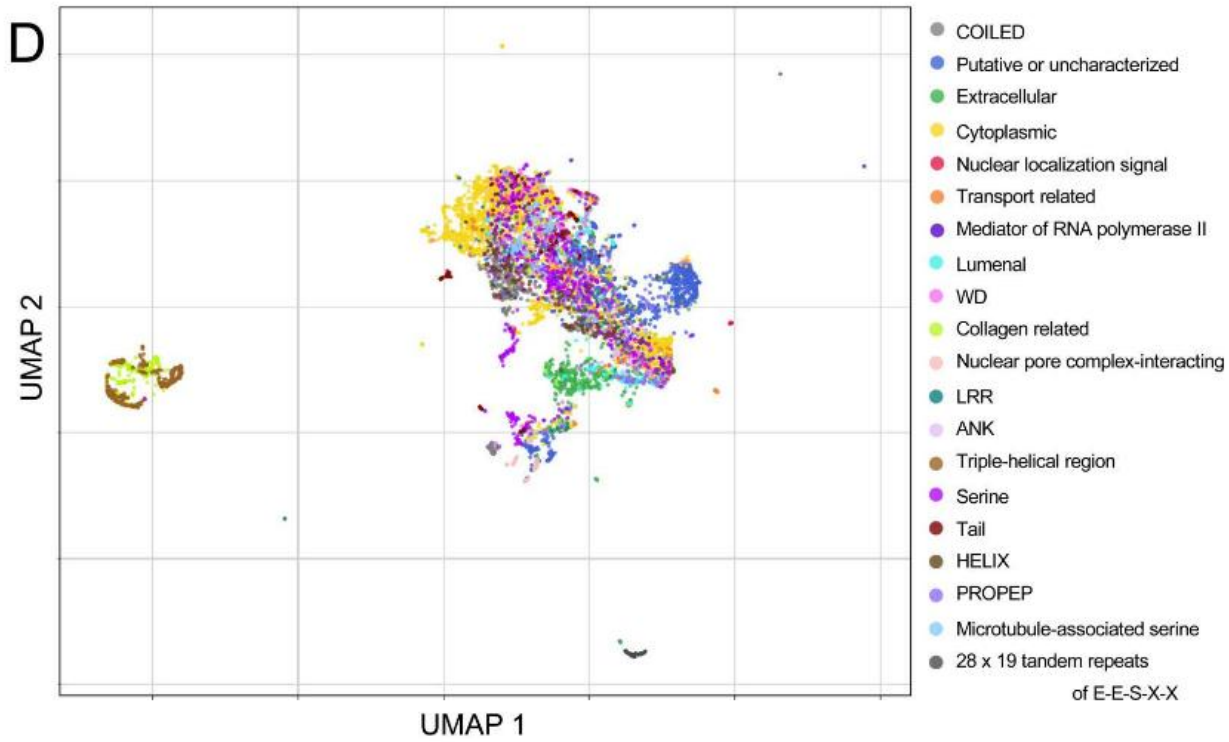




1-NN IDR Classification

IDR using Disprot and ProtGPS annotations

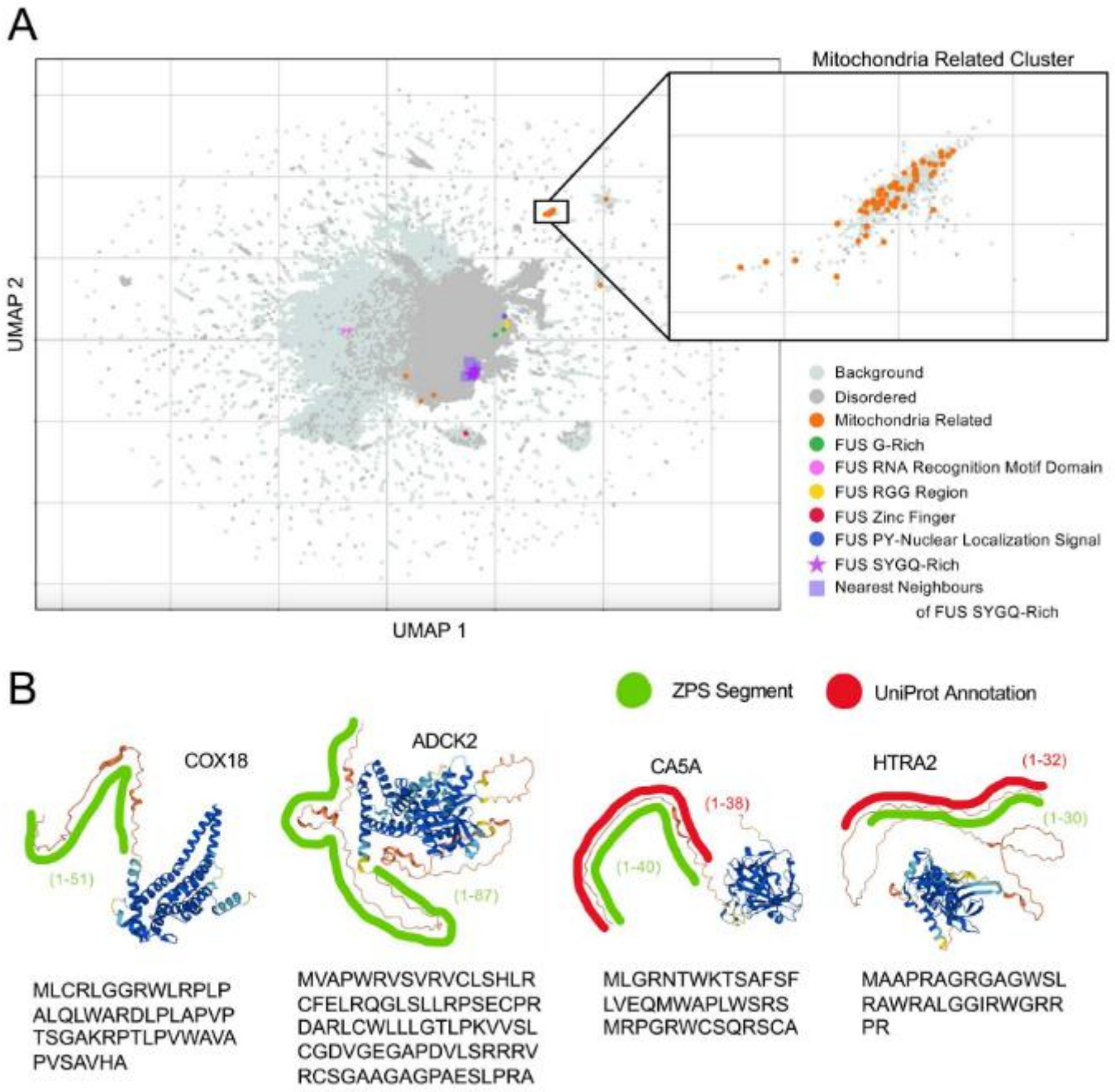




Discovery

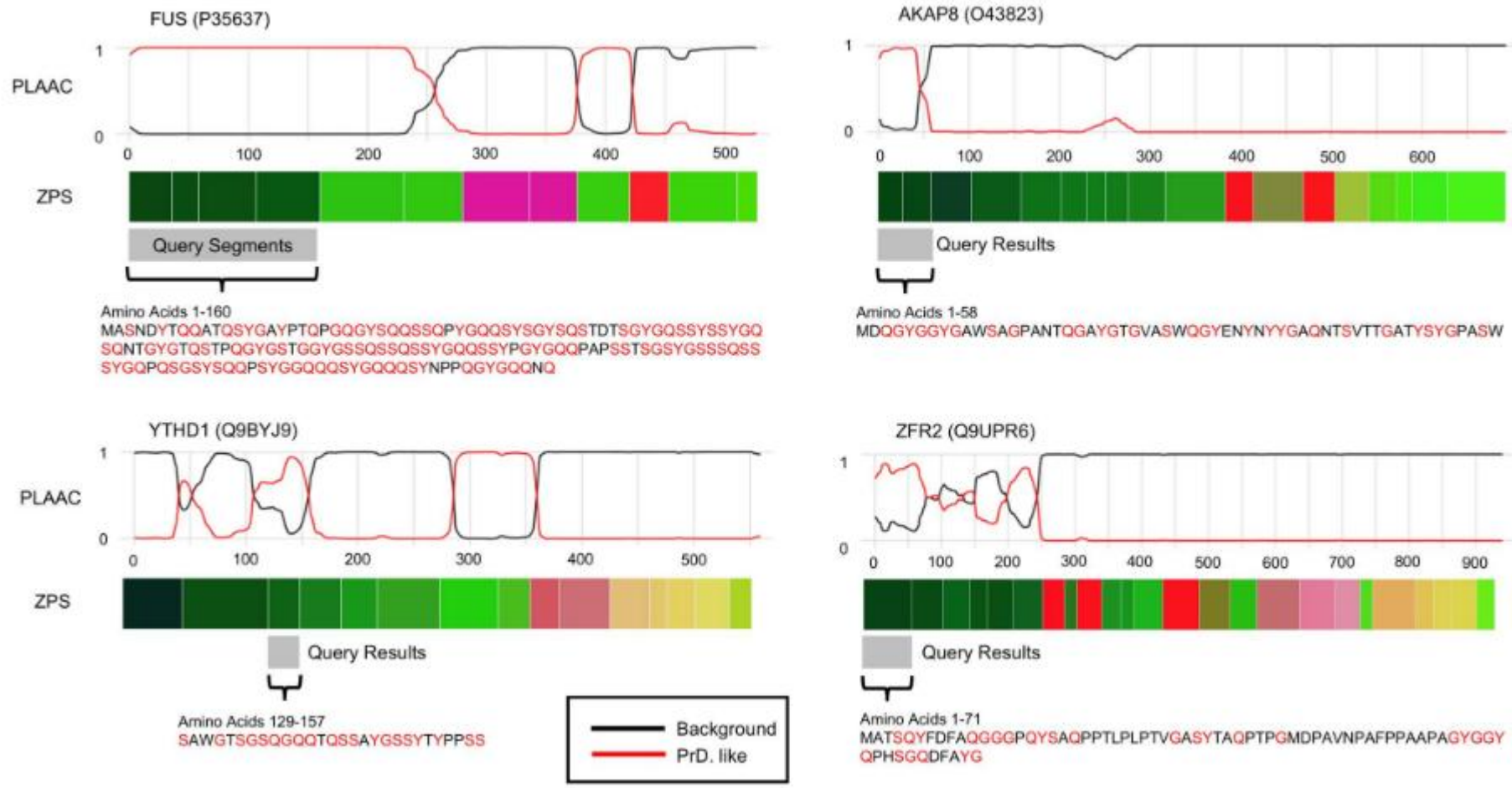
Mitochondria Related Cluster

defined by Leiden clustering of unannotated protein segments



Discovery

10 nearest-neighbours of FUS's SYGQ-Rich Region



Discussion

Strengths and Weaknesses of the paper

Strengths	Weaknesses
Segmentation is done with a simple zero shot method that *only uses sequence + PLM embeddings*	• **Over-segmentation** • **IoU** scores for segment evaluation do **not** show **significant improvement** over baselines
High accuracy for domain and IDR type 1NN based **classification**	• Classification done with **small set of labels** • Comparison with **baseline 3mer** embeddings only **shows strength of ProtT5 embeddings not segmentation** • 1-NN method allows **segments from the same protein (leakage)**
RGB colors of embeddings increase **interpretability**	The **colors are not linked to functionality**
Discovered unannotated segment clusters with **support from literature**	**Manual** exploration of embedding space was required to detect new annotations

Thanks…