SUPPLEMENTARY MATERIAL

A Datasets

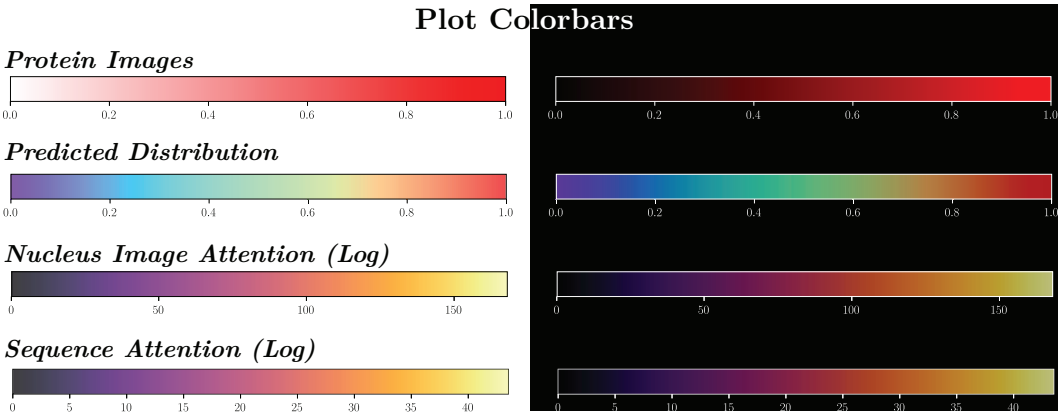

Figure S1: Colorbars used in figures on white (left) and black (right) background.

A.1 Human Protein Atlas

We used the Human Protein Atlas v21, available under the Creative Commons Attribution-ShareAlike 3.0 International License. For pre-training, we selected the immunofluorescence stained images from the Human Protein Atlas (HPA), which contains data on more than 17,268 human proteins, with information on their distribution across 44 different normal human tissues and 20 different cancer types. Example images show distribution of proteins within 2-5 cell types with different antibody markers [1]. We extracted corresponding amino acid sequences from UniProt [2].

A.2 OpenCell

We selected the OpenCell dataset for fine tuning due to its high-quality images, consistent imaging and cell conditions, and availability of reference images with consistent morphology. The dataset includes a collection of 1,311 CRISPR-edited HEK293T human cell lines, each tagged with a target protein using the split-mNeonGreen2 system. For each cell line, the OpenCell imaging dataset contains 4-5 confocal images of the tagged protein, accompanied by DNA staining to serve as a reference for nuclei morphology. While smaller in comparison to HPA, the cells were imaged while alive, providing a more accurate representation of protein distribution within the cell than immunofluorescence [3]. The OpenCell dataset is available under the BSD 3-Clause License.

A.3 Amino Acid Sequence Preprocessing

In natural language contexts, ensuring input sequences are the same length is usually performed by modifying the end of the sequence, either via truncation or end-padding [4]. This allows for predictions with respect to a given input (i.e. a text prompt). From the perspective of protein function, however, both the beginning and end (N and C termini) are points of interest for appending amino acids, especially with respect to protein localization [5, 6]. As such, we augment the sequence data as follows:

1. The amino acid sequence is tokenized using the ESM-2 tokenizer.
2. Start and end tokens are appended to the beginning and end of the sequence.
3. Cropping or padding occur based on the full sequence length, (length of amino acid sequence + `<START>` token + `<END>` token = 1002).
    - If the full sequence length $>$ 1002 tokens, we randomly crop 1002 tokens.
    - If the full sequence length $<$ 1002 tokens, we randomly add pad tokens before the `<START>` token and/or after the `<END>` token (See Fig. S2).

4. A <SEP> token is appended to the end of the protein sequence.

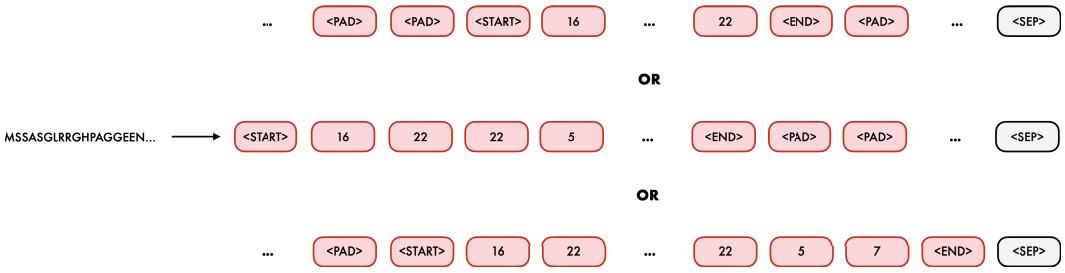

Figure S2: The amino acid sequence is tokenized and randomly padded via the <PAD> token. The top row shows start and end padding. The middle row shows end padding. The bottom row shows start-padding. All of these are possible. Note that the fixed length of 1002 means that the <SEP> token is always placed in the 1003rd position.

A.4 Image Preprocessing

A few preprocessing steps were necessary for the image encoder. Our image processing procedure is as follows:

1. We clip pixels beneath the .001 and above the 99.999 percentiles.
2. We normalize image values based on the calculated means and standard deviation from the datasets:
   Human Protein Atlas
   Nucleus: $\mu = 0.0655, \quad \sigma = 0.1732$
   Protein Image: $\mu = 0.0650, \quad \sigma = 0.1208$
   OpenCell
   Nucleus: $\mu = 0.0272, \quad \sigma = 0.0486$
   Protein Image: $\mu = 0.0244, \quad \sigma = 0.0671$
3. We rescale the images so pixel values are between 0 and 1.
4. The median pixel value of the protein image is calculated to create the thresholded image such that pixels $\geq$ median = 1 and pixels $<$ median $= 0$.
   Finally, we rescale images to 600×600 and randomly crop to 256×256 pixels.
5. Data augmentation is applied via random horizontal and vertical flips.

B Methods

B.1 Sampling

We experimented with the cosine-scheduling approach used in other works [7, 8], but we did not see any improvement in reconstruction performance (Fig. S4). We predicted the entire image in one step for image prediction. For amino acid sequence prediction, we predict amino acids one-by-one from the central protein.

We also calculated the probabilities of each token for all image predictions. We kept the output logits of the transformer. For image logits, we normalized them to 1 and fed them to the VQGAN decoder, which performed a linear interpolation in latent space. We clipped the values between 0 and 1 and displayed them as a heatmap (Fig. S3).

B.2 Training

We utilized 4$\times$ NVIDIA RTX 3090 TURBO 24G GPUs for this study. 2 GPUs were utilized for training VQGANs via distributed training. Our computer also contained 2$\times$ Intel Xeon Silver and

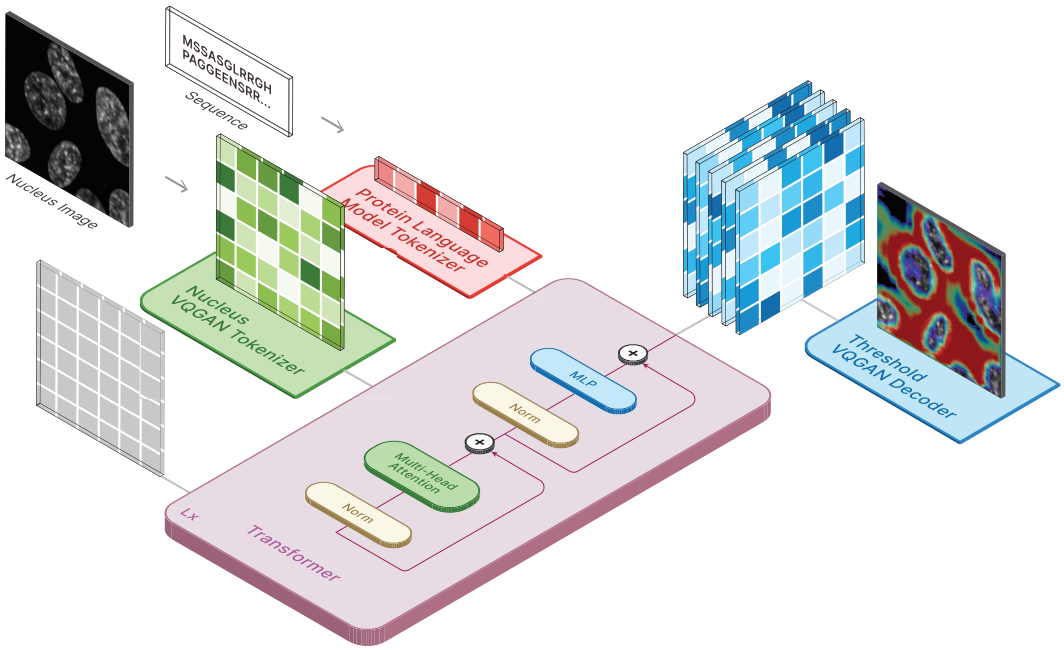

Figure S3: Depiction of the reconstruction scheme used to generate the predicted distribution heatmaps. Similar to training time, we provide tokenized vectors corresponding to the amino acid sequence and the nucleus image. Every position for the tokenized image is set to `<MASK_IM>` (shown as gray squares). The output logits are saved for every position and treated as probabilities associated with each image patch. These values are scaled and sent to the threshold VQGAN decoder to produce the final heatmap. Pixel values in the final image are clipped between 0 and 1.

8× 32768mb 2933MHz DR×4 Registered ECC DDR4 RAM. Only a single GPU is ever used to train CELL-E 2 models. Models were implemented in Python 3.11 using Pytorch 2.0 [9].

In order to train the transformer, we underwent the following procedure (Fig. 2):

1. We tokenize the amino acid sequence using the ESM-2 dictionary. We tokenize the nucleus image and protein threshold image using the codebook indices of the respective pre-trained VQGANs.
2. We retrieve embeddings for the amino acid sequence from the pre-trained ESM-2 protein language model (available under the MIT license Copyright (c) Meta Platforms, Inc. and affiliates.) . These embeddings are frozen and never updated over the course of training.
3. We randomly mask the amino acid sequence and protein threshold image tokens. The `<SEP>` and nucleus image tokens are never masked.
4. We obtain embeddings for the image tokens from embedding spaces created within the transformer and are learned over training. These size of the embedding are set to the same dimension as the pre-trained language embeddings. We similarly retrieve embeddings from a separate embedding space for the `<SEP>` token.
5. We pass the embeddings through a positional encoder via rotary encoding [10].
6. We concatenate the embeddings along the sequence dimension and pass them through the transformer. We calculate loss via cross-entropy only on the masked tokens.

Hyperparameters We used the following hyperparameters for our transformer model. Based on the findings of Khwaja et al. [11], we increased the transformer depth to achieve better predictive performance. The "Embedding Dimension" was determined by the protein language model we used, so we maximized the number of layers within the transformer, constrained by the VRAM capacity.

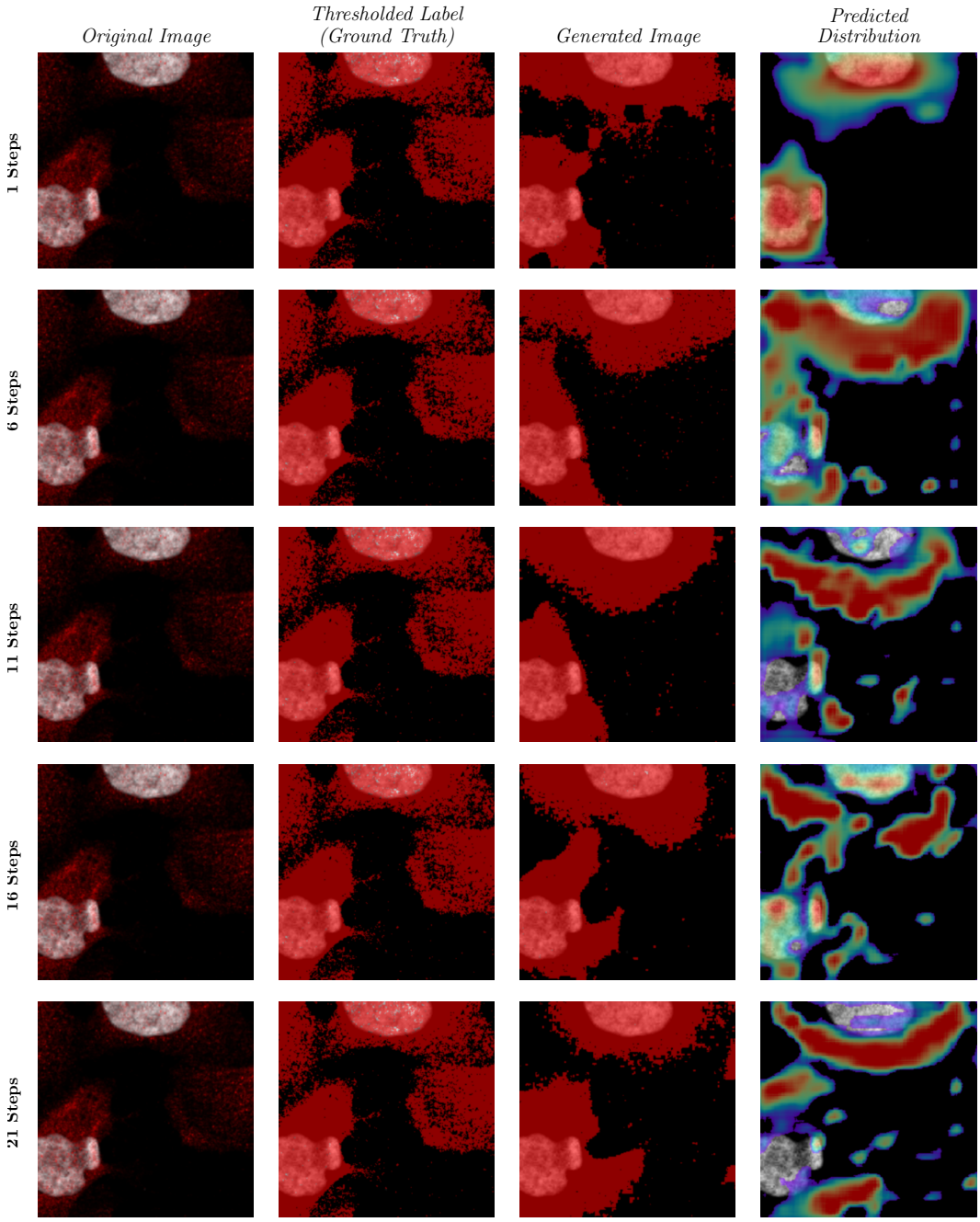

Figure S4: Image prediction based on the number of reconstruction steps. Note the decreased distribution intensity with increasing step count.

Table S1: VQGAN Hyperparameters

Hyperparameter	Value
Optimizer	Adam [12]
Base Learning Rate	4.5×10^{-6}
Betas	$\beta_1 = .5, \beta_2 = .9$
Weight Decay	0
Embedding Dimension	256
Number of Embeddings	512
Resolution	256
Number of Input Channels	1
Dropout	0
Discriminator Start	50000
Discriminator Weight	.2
Codebook Weight	1.0

Table S2: Base Transformer Hyperparameters

Hyperparameter	Value
Optimizer	AdamW [13]
Base Learning Rate	3×10^{-4}
Betas	$\beta_1 = .9, \beta_2 = .95$
Weight Decay	.01
Number of Text Tokens	33
Text Sequence Length	1000
Embedding Dimension/Depth	480/68 or 640/55 or 1280/25 or 2560/5
Number of Heads	16
Dimension of Head	64
Attention Dropout	.1
Feedforward Dropout	.1
Image Loss Weight	1
Condition Loss Weight	1

Table S3: CELL-E 2 Model Parameters per Size

Embedding Dimension/Depth	# of Params
480/68	536 M
640/55	744 M
1280/25	1.46 B
2560/5	3.47 B

C Results

C.1 Image Prediction Accuracy

Table S4 shows the image prediction performance of HPA and OpenCell-trained across both datasets and splits. We evaluate image reconstruction using the following metrics:

Nucleus Proportion MAPE This metric measures how well the predicted protein image matches the ground truth in terms of the fraction of intensity within the nucleus. We use Cellpose [14] to create a mask of the nucleus channel. Then we divide the sum of the predicted 2D PDF pixels inside the mask by the sum of all pixels in the image. We do the same for the ground truth protein image and compare the two fractions. The error is expressed as a percentage of the ground truth fraction.

Image MAE This metric calculates the average absolute difference between each pixel in the predicted protein threshold image and the ground truth protein threshold image. A lower MAE means a better match.

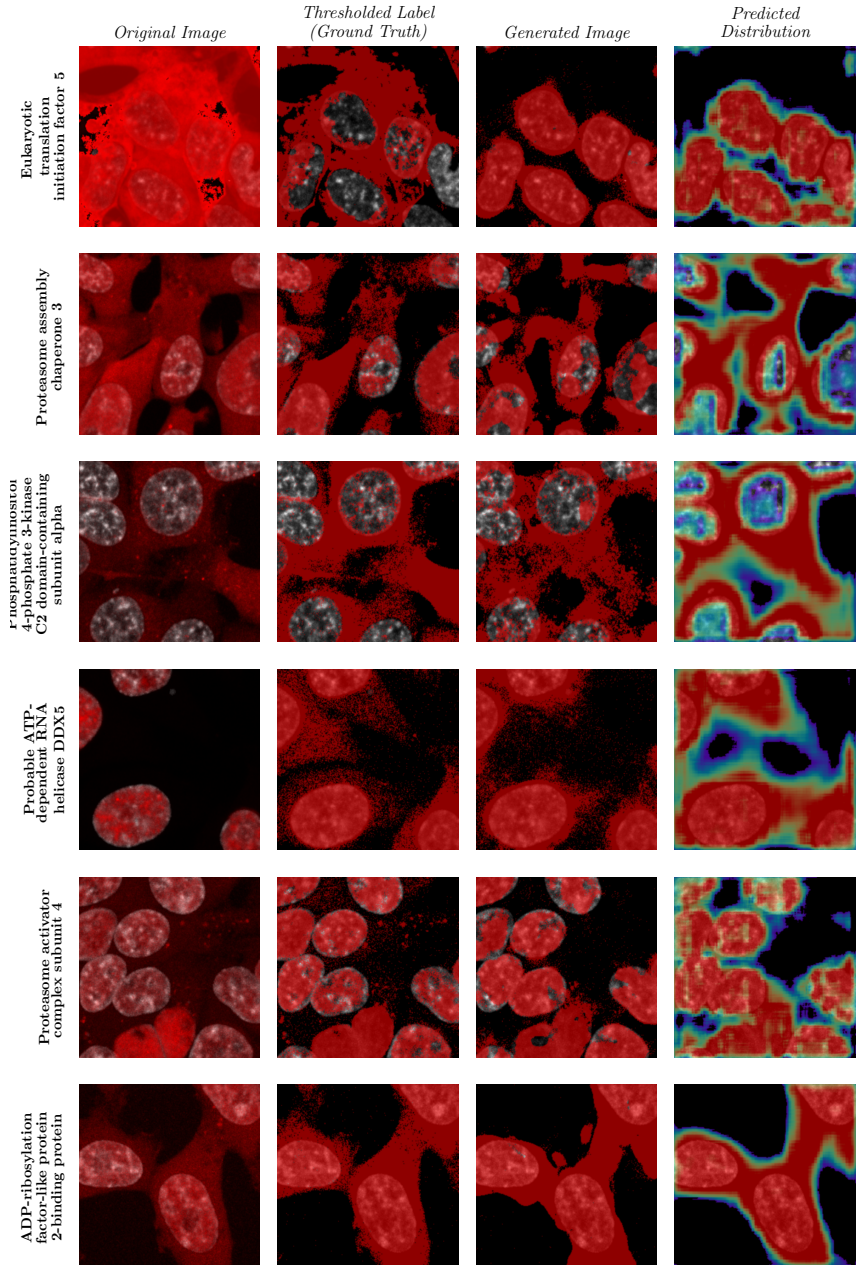

Figure S5: More randomly selected predictions from `HPA Finetuned HPA VQGAN_480`. We only note an incorrect prediction in Eukaryotic translation initiation factor 5.

PDF MAE This metric is similar to Image MAE, except we evaluate the difference using the predicted 2D PDF, rather than the predicted protein threshold image. We expect this number to be less accurate as tokens with less confidence will reduce the pixel value, while all values in the protein threshold image are 0 or 1.

SSIM Structural similarity index measure (SSIM) is a metric that evaluates how similar two images are in terms of local features such as brightness and contrast. It takes into account the spatial relationships between neighboring pixels. SSIM values range from 0, meaning no similarity, to 1, meaning perfect similarity.

Table S4: Image Prediction Accuracy Across OpenCell and HPA

Training Set Proteins

Dataset	Train Set	Hidden Size	Depth	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
HPA	HPA	480	68	.0254 ± .0296	.3344 ± .0797	.2845 ± .0991	.2635 ± .1797	11.4596	2.3151 ± .0224
		640	55	.0291 ± .0318	**.3286 ± .0808**	**.2843 ± .0996**	**.2827 ± .1836**	21.0591	2.2879 ± .0153
		1280	25	.0356 ± .0341	.3640 ± .0797	.2942 ± .0973	.2673 ± .1862	**1.0080**	2.5634 ± .0192
		2560	5	.0788 ± .0773	.3530 ± .0795	.3097 ± .0904	.2569 ± .1636	22.8721	2.1817 ± .0166
	OpenCell	480	68	**.0244 ± .0317**	.4620 ± .0769	.3530 ± .0803	.0865 ± .0714	4.1290	**2.7063 ± .0146**
		640	55	.0247 ± .0285	.4676 ± .0778	.3572 ± .0781	.0800 ± .0674	37.6196	2.4858 ± .0169
		1280	25	.0368 ± .0321	.4678 ± .0776	.3835 ± .0659	.0712 ± .0518	21.3462	1.5207 ± .0020
		2560	5	.0706 ± .0737	.4678 ± .0777	.3474 ± .0797	.1041 ± .0725	14.4177	1.7531 ± .0109
OpenCell	HPA	480	68	.0184 ± .0177	.4138 ± .0573	.3699 ± .1262	.1388 ± .1206	3.7217	2.3090 ± .0548
		640	55	.0183 ± .0166	**.4087 ± .0579**	.3835 ± .1191	.1230 ± .1128	3.5440	2.0354 ± .0998
		1280	25	.0219 ± .0202	.4358 ± .0588	.3659 ± .1141	.1225 ± .1198	7.1451	2.1888 ± .0776
		2560	5	.0460 ± .0418	.4164 ± .0693	.3905 ± .0962	.0984 ± .0870	7.5480	2.0104 ± .0519
	OpenCell	480	68	**.0134 ± .0131**	.4930 ± .0074	**.3264 ± .1108**	**.1620 ± .1429**	**.8923**	**3.0345 ± .1000**
		640	55	.0141 ± .0124	.4994 ± .0006	.3473 ± .0995	.1291 ± .1195	2.8314	2.3160 ± .0702
		1280	25	.0277 ± .0230	.4996 ± .0007	.4276 ± .0707	.0743 ± .0518	9.3420	1.3759 ± .0213
		2560	5	.0567 ± .0479	.4996 ± .0006	.4037 ± .0877	.0927 ± .0681	9.8328	1.4463 ± .0260

Validation Set Proteins

Dataset	Train Set	Hidden Size	Depth	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
HPA	HPA	480	68	.0257 ± .0250	.3340 ± .0788	.2846 ± .0985	.2633 ± .1781	12.0332	2.2900 ± .0410
		640	55	.0294 ± .0278	**.3283 ± .0805**	**.2842 ± .0991**	**.2826 ± .1827**	21.7942	2.2618 ± .0364
		1280	25	.0370 ± .0360	.3622 ± .0799	.2967 ± .0985	.2645 ± .1857	**1.5161**	2.5440 ± .0490
		2560	5	.0818 ± .0794	.3516 ± .0792	.3104 ± .0904	.2558 ± .1619	23.7977	2.1578 ± .0290
	OpenCell	480	68	**.0245 ± .0235**	.4622 ± .0767	.3533 ± .0803	.0861 ± .0718	41.5344	**2.6712 ± .0225**
		640	55	.0248 ± .0231	.4676 ± .0776	.3575 ± .0783	.0795 ± .0681	38.3386	2.4850 ± .0381
		1280	25	.0371 ± .0343	.4678 ± .0775	.3833 ± .0661	.0713 ± .0525	21.6973	1.5206 ± .0152
		2560	5	.0717 ± .0722	.4678 ± .0776	.3474 ± .0796	.1038 ± .0731	14.7231	1.7524 ± .0160
OpenCell	HPA	480	68	.0181 ± .0168	.4154 ± .0594	.3887 ± .1270	.1250 ± .1149	3.9509	2.1739 ± .1255
		640	55	.0178 ± .0165	**.4058 ± .0574**	.3651 ± .1197	**.1359 ± .1183**	3.0867	2.1508 ± .0384
		1280	25	.0227 ± .0213	.4323 ± .0581	.3886 ± .1128	.1051 ± .1140	**1.4713**	2.0247 ± .1003
		2560	5	.0487 ± .0453	.4202 ± .0722	.4049 ± .0870	.0874 ± .0792	9.1799	1.9269 ± .0768
	OpenCell	480	68	.0161 ± .0148	.4953 ± .0064	**.3620 ± .1168**	.1220 ± .1188	1.5844	**2.6069 ± .1175**
		640	55	**.0159 ± .0136**	.4995 ± .0006	.3785 ± .1008	.1011 ± .1012	2.6966	2.0974 ± .0981
		1280	25	.0272 ± .0223	.4996 ± .0010	.4359 ± .0700	.0694 ± .0472	8.9102	1.3712 ± .0432
		2560	5	.0584 ± .0511	.4996 ± .0005	.4145 ± .0889	.0890 ± .0667	9.5116	1.4176 ± .0329

IS Inception score (IS) is a metric that assesses how realistic and diverse the images generated by a model are. It uses a pretrained neural network to classify the images and computes a score based on how well they fit into different categories. A higher IS means more realistic and varied images.

FID Fréchet Inception Distance (FID) is another metric that compares the quality and diversity of generated images to ground truth images. It calculates the distance between two statistical representations of the image distributions, called feature vectors, which are extracted by a pretrained neural network. A lower FID means more similar distributions and better quality images. For this study FID was scored against the training or validation sets when applicable.

C.2 Masked Sequence In-Filling

Table S6 shows the sequence prediction performance (predicting 15% of masked residues) of the models shown in Table S4. We evaluate only on masked positions using the following criteria:

Sequence MAE This metric calculates the average absolute difference between each amino acid in the predicted sequence and the ground truth sequence for the masked positions. A lower MAE means a better match.

Cosine Similarity We evaluate cosine similarity of the amino acid embeddings. This metric measures the angle between two vectors that represent the predicted sequence and the ground truth

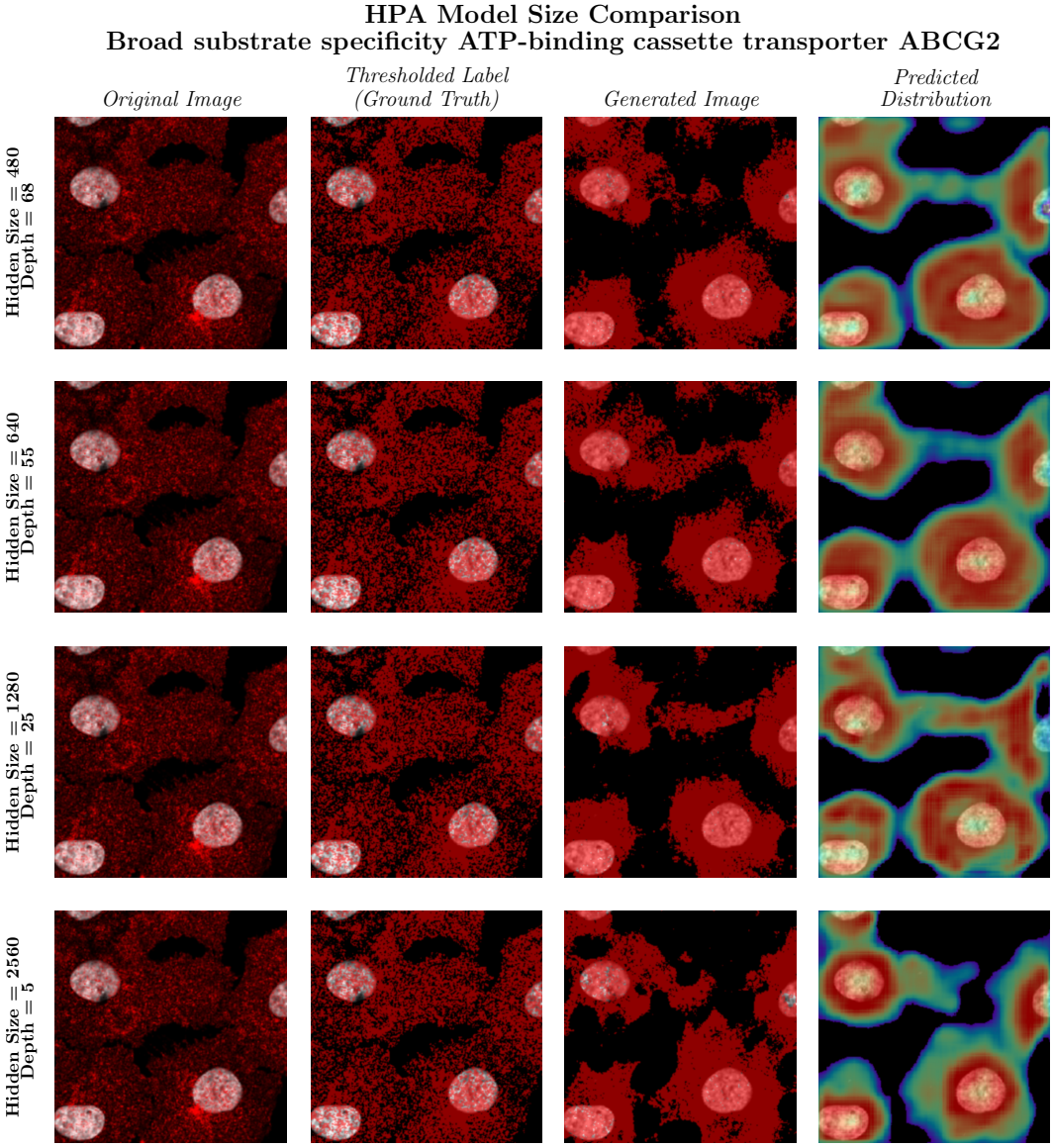

Figure S6: CELL-E 2 models trained on the HPA dataset. Predictions are shown based on the hidden size of the transformer embedding. We see the strongest performance from the 480 and 640 models. Localization is expected within the mitochondria in the selected protein. Not the heightened intensity within the nuclear region in the 1280 and 2560 models predictions.

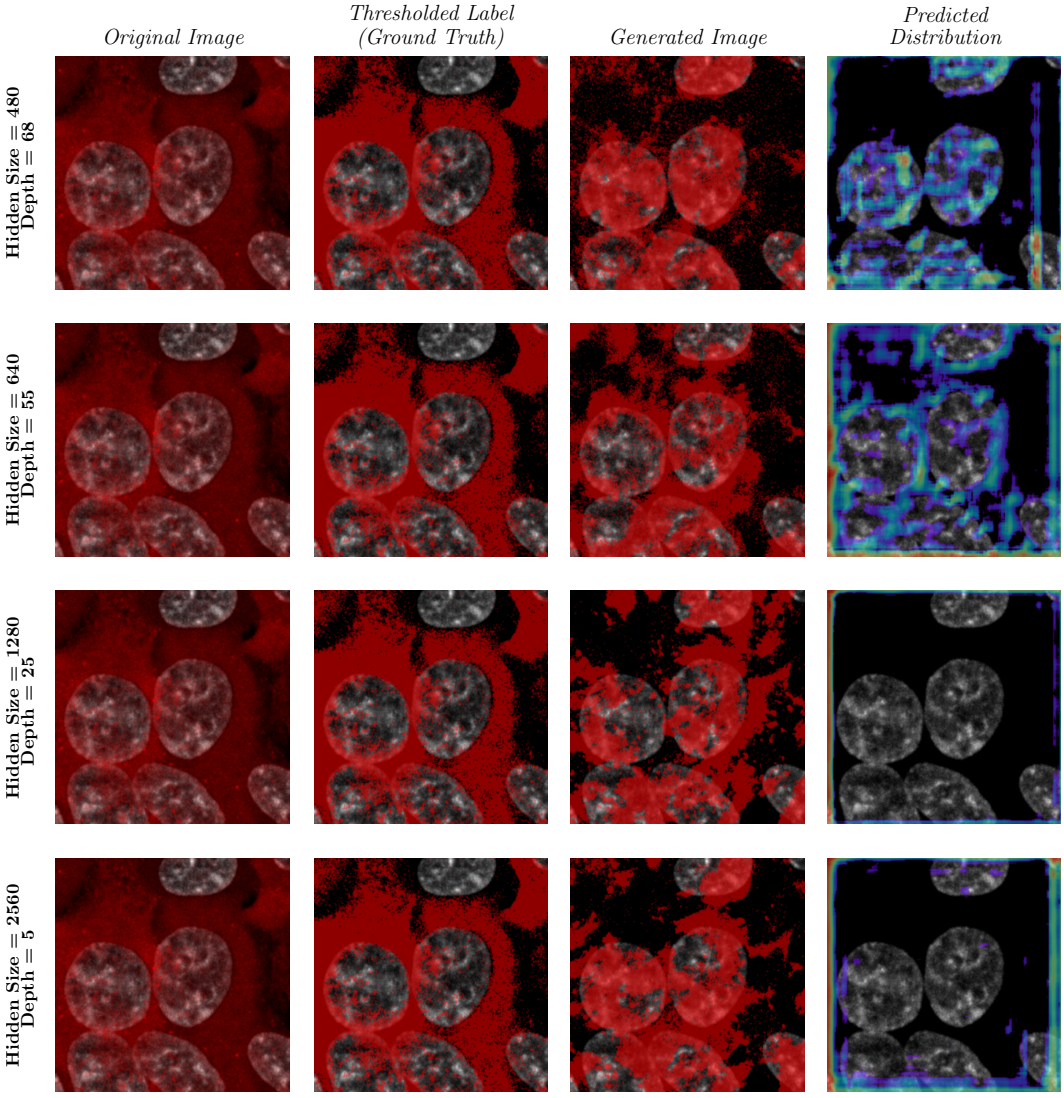

Figure S7: Similar to Fig. S6, we depict the performance of CELL-E 2 models only trained on the OpenCell dataset. We see the best performance on the 480 model, but not drastically different predicted distribution images. This is likely a function of reduced training time due to the quick overfitting of the model.

sequence. It ranges from -1 to 1, where 1 means the vectors are identical, 0 means they are orthogonal, and -1 means they are opposite. A higher cosine similarity means a more similar sequence. Note that cosine similarity is performed on the entirety of the protein and not just masked positions.

C.3 Finetuning

Table S8 shows the image prediction performance of models across datasets after fine-tuning on the OpenCell dataset. Table S9 shows the sequence prediction accuracy of the same models.

Table S5: ESM-2 Masked Sequence In-Filling Accuracy (No Image)

Training Set Proteins

Dataset	Hidden Size	# Layers	Sequence MAE	Cosine Similarity
HPA	480	12	.7351 ± .1100	.9464 ± .0232
	640	30	.6507 ± .1317	.9572 ± .0183
	1280	33	.4921 ± .1741	.9724 ± .0133
	2560	36	**.3818 ± .1911**	**.9778 ± .0130**
OpenCell	480	12	.7276 ± .1144	.9425 ± .0233
	640	30	.6151 ± .1364	.9572 ± .0159
	1280	33	.4335 ± .1650	.9746 ± .0082
	2560	36	**.3298 ± .1762**	**.9793 ± .0089**

Validation Set Proteins

Dataset	Hidden Size	# Layers	Sequence MAE	Cosine Similarity
HPA	480	12	.7368 ± .1116	.9471 ± .0209
	640	30	.6553 ± .1334	.9571 ± .0161
	1280	33	.5005 ± .1705	.9723 ± .0096
	2560	36	**.3894 ± .1911**	**.9777 ± .0096**
OpenCell	480	12	.7355 ± .1130	.9381 ± .0286
	640	30	.6185 ± .1454	.9538 ± .0199
	1280	33	.4260 ± .1822	.9737 ± .0096
	2560	36	**.3220 ± .1848**	**.9789 ± .0086**

D Discussion

D.1 CELL-E Comparison

Table S10 shows the image prediction metrics for the original CELL-E model on both the HPA and OpenCell datasets. Note that CELL-E was only trained on OpenCell data.

Table S11 depicts the mean time taken for 10 separate model predictions. CELL-E is not directly comparable to CELL-E 2 due to differences in language model and package versioning, so we opt to include the compute time of CELL-E 2 using an autoregressive reconstruction scheme (i.e. 256 sequential steps from top left to bottom right). CELL-E 2 model run in autoregressive mode are significantly slower due to the lack of cache implementation found in CELL-E and the larger ESM-2 language model compared to the TAPE model used in CELL-E. CELL-E 2 models which generate the prediction in a single step (NAR) are an orders of magnitude faster than their autoregressive counterparts.

D.2 De novo NLS Design

NLS generation

1. We selected a desired NLS length (iterating over a range of 5 to 30 residues) and inserted that number of mask tokens after the starting methionine in the GFP sequence. (e.g. an NLS of length 5 at the N terminus would have an input sequence of `<START> M <MASK_SEQ> <MASK_SEQ> <MASK_SEQ> <MASK_SEQ> <MASK_SEQ>SKGEE...<END> <PAD>...`).
2. We randomly chose a nucleus image and segmented the nuclei area by applying a mask with Cellpose [14]. We assigned the pixels inside the nucleus area to True and used this as the threshold image.
3. We inputted the masked GFP sequence, the nucleus image, and the threshold image to the transformer and sampled the output. We used the model depth that achieved the highest performance on sequence reconstruction, which was `OpenCell_2560`.

Table S6: Masked Sequence In-Filling Accuracy

Training Set Proteins

Dataset	Train Set	Hidden Size	Depth	Sequence MAE	Cosine Similarity
HPA	HPA	480	68	.8548 ± .1050	.9500 ± .0260
		640	55	.7738 ± .1368	.9580 ± .0238
		1280	25	.5818 ± .2053	**.9733 ± .0195**
		2560	5	**.5294 ± .2402**	.9732 ± .0235
	OpenCell	480	68	.8554 ± .1047	.9504 ± .0262
		640	55	.7806 ± .1343	.9576 ± .0239
		1280	25	.6377 ± .1850	.9709 ± .0191
		2560	5	.5599 ± .2294	.9721 ± .0235
OpenCell	HPA	480	68	.8403 ± .1102	.9463 ± .0277
		640	55	.7434 ± .1356	.9557 ± .0263
		1280	25	.5315 ± .1996	.9725 ± .0219
		2560	5	.4760 ± .2281	**.9726 ± .0266**
	OpenCell	480	68	.7507 ± .1709	.9533 ± .0285
		640	55	.6641 ± .1764	.9610 ± .0272
		1280	25	.5698 ± .2016	.9709 ± .0220
		2560	5	**.4950 ± .2456**	.9711 ± .0271

Validation Set Proteins

Dataset	Train Set	Hidden Size	Depth	Sequence MAE	Cosine Similarity
HPA	HPA	480	68	.8628 ± .0951	.9504 ± .0237
		640	55	.7917 ± .1245	.9577 ± .0216
		1280	25	.6512 ± .1794	.9708 ± .0163
		2560	5	.5759 ± .2322	.9722 ± .0210
	OpenCell	480	68	.8625 ± .0935	.9508 ± .0240
		640	55	.7927 ± .1245	.9577 ± .0216
		1280	25	.6476 ± .1811	.9711 ± .0163
		2560	5	**.5696 ± .2288**	**.9724 ± .0210**
OpenCell	HPA	480	68	.8651 ± .0992	.9420 ± .0312
		640	55	.7675 ± .1318	.9529 ± .0271
		1280	25	.5910 ± .2065	.9699 ± .0213
		2560	5	.5137 ± .2414	.9700 ± .0250
	OpenCell	480	68	.8600 ± .1030	.9430 ± .0316
		640	55	.7645 ± .1332	.9532 ± .0273
		1280	25	.5872 ± .2060	.9703 ± .0213
		2560	5	**.5080 ± .2365**	**.9703 ± .0250**

4. For each sequence length, we generated 300 candidates per length per terminus. We then provided the `HPA Finetuned (Finetuned HPA VQGAN)_480` model with the predicted NLS + GFP sequence and the nucleus image. Using the previously calculated nucleus mask, we calculate the percentage of positive intensity predicted within the nucleus bounds. Any sequence with a predicted nucleus proportion intensity $< 75\%$ was discarded.

We generated candidate NLS with lengths from 2 to 30 amino acids at the N and C termini of the protein. We ranked them using these criteria:

- **Forward Consistency:** The proportion of positive signal in the nucleus mask relative to the whole image, using the best image prediction model (480 model), similar to Section 5.1.

Table S7: Masked Sequence Random In-Filling Accuracy

Training Set Proteins

Dataset	Sequence MAE	Cosine Similarity
HPA	.9600 ± .0274	.9502 ± .0181
OpenCell	.9603 ± .0268	.9469 ± .0176

Validation Set Proteins

Dataset	Sequence MAE	Cosine Similarity
HPA	.9605 ± .0257	.9509 ± .0157
OpenCell	.9592 ± .0282	.9461 ± .0191

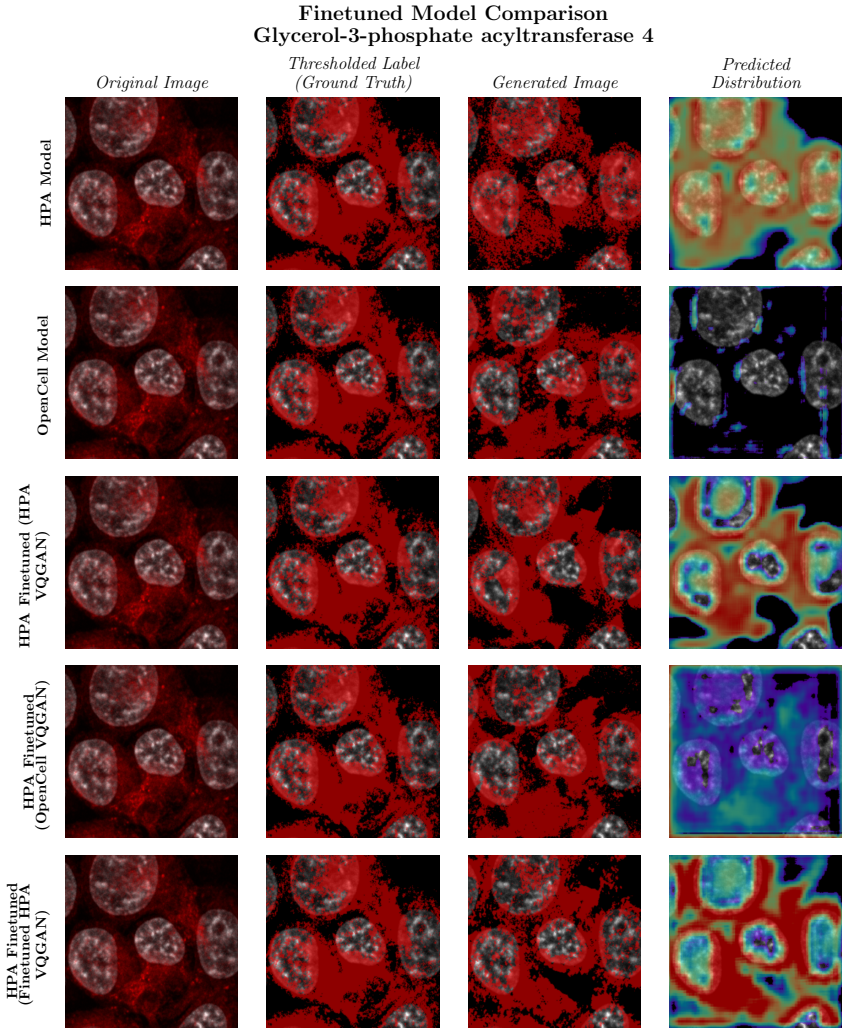

Figure S8: Various model performance from different fine tuning methods. We note superior predictive performance from the model with where we initially fine-tune the image encoder.

- **Image Prediction Confidence:** The values from the predicted distribution using a masked approach, indicating the confidence in the localization image prediction.

Table S8: Image Prediction Accuracy after Finetuning on HPA and OpenCell

Training Set Proteins

Dataset	Image Encoders	Hidden Size	Depth	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
HPA	HPA	480	68	.0292 ± .0291	.3606 ± .0832	.3599 ± .0836	**.2237 ± .1479**	22.0947	2.8130 ± .0208
	Opencell			**.0245 ± .0317**	.4680 ± .0776	.3428 ± .0833	.1047 ± .0840	23.2398	3.0922 ± .0167
	HPA Finetuned			.0249 ± .0289	.3755 ± .1011	**.3292 ± .0848**	.1406 ± .1027	8.3675	**3.9647 ± .0299**
	HPA	640	55	.0299 ± .0263	.3475 ± .0834	.3472 ± .0819	.1516 ± .1118	6.7563	2.0455 ± .0099
	Opencell			.0273 ± .0254	.4518 ± .0570	.3505 ± .0778	.0900 ± .0747	31.7937	2.5763 ± .0119
	HPA Finetuned			.0270 ± .0249	**.3041 ± .0907**	.3328 ± .0794	.1278 ± .0910	11.4788	2.3392 ± .0130
	HPA	1280	25	.0448 ± .0400	.3461 ± .0820	.3350 ± .0842	.2004 ± .1364	6.8770	2.1677 ± .0096
	OpenCell			.0426 ± .0410	.4486 ± .0556	.3401 ± .0826	.1067 ± .0841	17.6565	2.7158 ± .0105
	HPA Finetuned			.0435 ± .0437	.3315 ± .0888	.3323 ± .0826	.1762 ± .1183	**5.9633**	2.2360 ± .0279
	HPA	2560	5	.0729 ± .0655	.3844 ± .0704	.3590 ± .0792	.1793 ± .1161	12.6113	2.0646 ± .0112
	OpenCell			.0727 ± .0776	.4736 ± .0633	.3428 ± .0847	.1291 ± .0925	8.4963	2.1803 ± .0116
	HPA Finetuned			.0744 ± .0671	.3507 ± .0803	.3599 ± .0795	.2014 ± .1322	16.672	2.2908 ± .0156
OpenCell	HPA	480	68	.0157 ± .0151	.3712 ± .0791	.3699 ± .0799	.2038 ± .1525	17.1616	3.0822 ± .0843
	OpenCell			**.0135 ± .0135**	.4996 ± .0007	**.3161 ± .1117**	.1874 ± .1495	1.5167	**3.0898 ± .1459**
	HPA Finetuned			.0154 ± .0150	**.3170 ± .1159**	.3186 ± .1215	**.2125 ± .1600**	18.7426	3.9276 ± .1406
	HPA	640	55	.0165 ± .0151	.4011 ± .0667	.3439 ± .1026	.1263 ± .1063	6.0163	2.2918 ± .0533
	OpenCell			.0149 ± .0136	.4732 ± .0192	.3415 ± .1054	.1356 ± .1281	4.9600	2.4016 ± .0866
	HPA Finetuned			.0167 ± .0150	.3305 ± .1035	.3400 ± .1059	.1525 ± .1195	2.8065	2.7464 ± .0621
	HPA	1280	25	.0243 ± .0224	.3817 ± .0686	.3355 ± .1065	.1546 ± .1201	3.7530	2.5043 ± .0454
	OpenCell			.0220 ± .0205	.4671 ± .0278	.3236 ± .1089	.1702 ± .1491	**.5084**	3.0222 ± .1054
	HPA Finetuned			.0254 ± .0241	.3701 ± .0838	.3581 ± .1054	.1468 ± .1156	5.2415	2.5990 ± .1403
	HPA	2560	5	.0411 ± .0379	.4067 ± .0745	.3363 ± .1087	.1775 ± .1299	14.7029	2.4132 ± .0603
	OpenCell			.0540 ± .0492	.4977 ± .0124	.3753 ± .1089	.1630 ± .1200	26.8886	1.8080 ± .0489
	HPA Finetuned			.0394 ± .0359	.3710 ± .0843	.3492 ± .1032	.1727 ± .1265	15.3433	2.5426 ± .0637

Validation Set Proteins

Dataset	Image Encoders	Hidden Size	Depth	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
HPA	HPA	480	68	.0291 ± .0259	.3589 ± .0838	.3583 ± .0843	**.2246 ± .1501**	21.8254	2.8176 ± .0210
	Opencell			**.0245 ± .0233**	.4681 ± .0774	.3430 ± .0833	.1047 ± .0853	23.9367	3.0918 ± .0519
	HPA Finetuned			.0249 ± .0235	.3427 ± .0908	**.3292 ± .0847**	.1397 ± .1047	8.7002	**3.9302 ± .0716**
	HPA	640	55	.0304 ± .0273	.3469 ± .0835	.3476 ± .0821	.1496 ± .1117	7.0875	2.0259 ± .0310
	Opencell			.0276 ± .0265	.4519 ± .0567	.3502 ± .0779	.0905 ± .0759	31.8870	2.5738 ± .0402
	HPA Finetuned			.0279 ± .0262	**.3041 ± .0906**	.3326 ± .0793	.1266 ± .0917	12.0062	2.3105 ± .0310
	HPA	1280	25	.0454 ± .0434	.3462 ± .0822	.3362 ± .0847	.1984 ± .1368	6.8893	2.1656 ± .0288
	Opencell			.0433 ± .0444	.4484 ± .0560	.3400 ± .0827	.1064 ± .0848	18.1654	2.7017 ± .0460
	HPA Finetuned			.0430 ± .0403	.3322 ± .0882	.3320 ± .0824	.1771 ± .1162	**5.9752**	2.2687 ± .0112
	HPA	2560	5	.0746 ± .0686	.3828 ± .0708	.3594 ± .0807	.1790 ± .1176	12.6199	2.0311 ± .0311
	OpenCell			.0739 ± .0755	.4730 ± .0650	.3429 ± .0854	.1289 ± .0957	8.7266	2.1980 ± .0275
	HPA Finetuned			.0761 ± .0697	.3510 ± .0816	.3603 ± .0810	.2003 ± .1332	16.4098	2.2785 ± .0319
OpenCell	HPA	480	68	.0166 ± .0151	.3776 ± .0834	.3477 ± .1268	**.1869 ± .1503**	17.4075	2.9113 ± .1199
	OpenCell			**.0159 ± .0156**	.4996 ± .0006	.3506 ± .1208	.1574 ± .1372	2.5026	2.7168 ± .1137
	HPA Finetuned			.0170 ± .0160	**.3449 ± .1305**	.3487 ± .1340	.1881 ± .1541	19.2683	**3.6083 ± .2013**
	HPA	640	55	.0176 ± .0155	.4028 ± .0668	.3644 ± .1004	.1060 ± .0928	7.9330	2.0560 ± .1219
	OpenCell			.0170 ± .0149	.4771 ± .0201	.3684 ± .1073	.1081 ± .1121	5.1479	2.1141 ± .1304
	HPA Finetuned			.0172 ± .0151	.3477 ± .1043	.3583 ± .1033	.1339 ± .1083	2.4811	2.4813 ± .1009
	HPA	1280	25	.0258 ± .0243	.3890 ± .0709	.3572 ± .1050	.1355 ± .1092	3.7844	2.2680 ± .1109
	OpenCell			.0262 ± .0259	.4743 ± .0275	.3576 ± .1133	.1339 ± .1218	**.9963**	2.6376 ± .1468
	HPA Finetuned			.0247 ± .0234	.3599 ± .0813	**.3361 ± .1078**	.1645 ± .1229	4.8118	2.8837 ± .0426
	HPA	2560	5	.0464 ± .0464	.4081 ± .0776	.3591 ± .1074	.1598 ± .1211	13.7206	2.2251 ± .1164
	OpenCell			.0594 ± .0533	.4969 ± .0121	.3928 ± .1074	.1509 ± .1135	27.7841	1.7532 ± .0837
	HPA Finetuned			.0446 ± .0430	.3812 ± .0885	.3709 ± .0988	.1549 ± .1193	13.4599	2.3191 ± .1147

- **Text Prediction Confidence:** The average probability values of the predicted NLS sequence tokens.
- **Sequence Similarity:** The maximum alignment score between the candidate NLS and sequences from the NLSdb, similar to Madani et. al. [15].
- **Embedding Cosine Angle:** The minimum cosine angle between the embeddings of the candidate NLS and sequences from the NLdb [16], using the same language model from Section 5.2, except similarity is evaluated on the entire protein sequence (NLS + GFP), rather than limited to the masked positions.

We rounded all values to one decimal place and ranked them by 1) Sequence Similarity, 2) Embedding Cosine Similarity, 3) Forward Consistency, 4) Image Prediction Confidence, 5) Text Prediction Confidence.

Table S9: Masked Sequence In-Filling Accuracy after Finetuning on HPA and OpenCell

Training Set Proteins

Dataset	Image Encoders	Hidden Size	Depth	Sequence MAE	Cosine Similarity
HPA	HPA	480	68	.8457 ± .1102	.9507 ± .0260
	OpenCell			.8442 ± .1144	.9508 ± .0259
	HPA Finetuned			.8498 ± .1108	.9506 ± .0259
	HPA	640	55	.7716 ± .1365	.9581 ± .0239
	OpenCell			.7729 ± .1422	.9582 ± .0240
	HPA Finetuned			.7755 ± .1354	.9579 ± .0239
	HPA	1280	25	.5742 ± .2022	**.9740 ± .0194**
	OpenCell			.5737 ± .2155	.9738 ± .0196
	HPA Finetuned			.5791 ± .2071	.9736 ± .0196
	HPA	2560	5	.5156 ± .2443	.9738 ± .0235
	OpenCell			.5177 ± .2426	.9736 ± .0236
	HPA Finetuned			**.5128 ± .2433**	.9739 ± .0236
OpenCell	HPA	480	68	.8139 ± .1436	.9483 ± .0279
	OpenCell			.7493 ± .1909	.9528 ± .0286
	HPA Finetuned			.8026 ± .1585	.9493 ± .0281
	HPA	640	55	.7339 ± .1560	.9560 ± .0267
	OpenCell			.6738 ± .1964	.9599 ± .0277
	HPA Finetuned			.7338 ± .1565	.9561 ± .0267
	HPA	1280	25	.4991 ± .2176	.9738 ± .0226
	OpenCell			.3697 ± .2493	**.9790 ± .0236**
	HPA Finetuned			.4959 ± .2190	.9740 ± .0229
	HPA	2560	5	.4510 ± .2568	.9725 ± .0273
	OpenCell			**.4289 ± .2600**	.9732 ± .0274
	HPA Finetuned			.4482 ± .2558	.9726 ± .0273

Validation Set Proteins

Dataset	Image Encoders	Hidden Size	Depth	Sequence MAE	Cosine Similarity
HPA	HPA	480	68	.8566 ± .1000	.9508 ± .0238
	OpenCell			.8575 ± .0973	.9507 ± .0237
	HPA Finetuned			.8610 ± .0998	.9507 ± .0238
	HPA	640	55	.7920 ± .1249	.9576 ± .0217
	OpenCell			.7976 ± .1243	.9574 ± .0217
	HPA Finetuned			.7954 ± .1235	.9575 ± .0216
	HPA	1280	25	.6434 ± .1840	.9713 ± .0163
	OpenCell			.6446 ± .1824	.9712 ± .0163
	HPA Finetuned			.5672 ± .2345	.9726 ± .0209
	OpenCell	2560	5	.5731 ± .2313	.9723 ± .0209
	HPA Finetuned			**.5651 ± .2329**	**.9727 ± .0210**
OpenCell	HPA	480	68	.8560 ± .1061	.9426 ± .0312
	OpenCell			.8634 ± .1101	.9421 ± .0313
	HPA Finetuned			.8689 ± .1090	.9417 ± .0311
	HPA	640	55	.7679 ± .1340	.9529 ± .0271
	OpenCell			.7829 ± .1385	.9517 ± .0276
	HPA Finetuned			.7792 ± .1398	.9520 ± .0273
	HPA	1280	25	.5955 ± .2134	.9695 ± .0218
	OpenCell			.5867 ± .2172	**.9698 ± .0219**
	HPA Finetuned			.5931 ± .2136	.9696 ± .0217
	HPA	2560	5	.5277 ± .2565	.9686 ± .0255
	OpenCell			.5322 ± .2545	.9684 ± .0255
	HPA Finetuned			**.5255 ± .2552**	.9687 ± .0255

Table S10: Image Prediction Accuracy for CELL-E

Training Set Proteins

Dataset	Hidden Size	Depth	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
HPA	768	32	.0672 ± .0632	.3601 ± .0829	.3303 ± .0796	.2219 ± .1383	4.2355	2.1292 ± .0139
OpenCell			.0377 ± .0327	.3642 ± .1150	.3600 ± .1044	.2133 ± .1825	11.3911	2.4390 ± .0625

Validation Set Proteins

Dataset	Hidden Size	Depth	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
HPA	768	32	.0786 ± .0644	.3610 ± .0816	.3308 ± .0785	.2217 ± .1371	4.1685	2.1021 ± .0371
OpenCell			.0347 ± .0294	.3671 ± .1117	.3653 ± .1008	.2060 ± .1846	10.5555	2.4762 ± .0866

Table S11: Speed Comparison

Model	Hidden Size	Autoregressive	Mean Generation Time (s)
CELL-E (Cached)	768	Yes	18.2740 ± 0.0451
CELL-E (Non-Cached)	768	Yes	28.7694 ± 0.3207
CELL-E 2	480	Yes	55.0057 ± 0.2069
CELL-E 2	640	Yes	62.9650 ± 0.1033
CELL-E 2	1280	Yes	74.3698 ± 0.1788
CELL-E 2	2560	Yes	128.9960 ± 0.3718
CELL-E 2	**480**	**No**	$\mathbf{0.2784 \pm 0.0006}$
CELL-E 2	640	No	0.3067 ± 0.0012
CELL-E 2	1280	No	0.3249 ± 0.0011
CELL-E 2	2560	No	0.5487 ± 0.0022

Table S12: NLS Composition

Max ID %	# Sequences	Mean Sequence Length	Mean % R or K
0% - 33%	109	25.6606 ± 3.0099	20.6379 ± 8.6101
33% - 66%	133	17.1955 ± 5.0804	32.0076 ± 12.8334
66% - 100%	13	6.9231 ± 1.2558	57.5794 ± 17.9351

D.3 Visualizing Attention

In Fig. S11 and Fig. 3, we depict the relative attention weights placed on the input amino acid sequence and nucleus image used to generate the threshold prediction. Specifically, we sought to emphasize weights correlated with positive signal, that is patches with largely white pixels. In this way, we do not bias the weights we consider with the use of any manual feature annotations or image segmentation. We first use attention rollout [17] to obtain the relative correlation between tokens at the end of the network. We then take an average across the multiplied attention heads. From here, we separate "positive" vs "negative" signal image patches based on the average intensity within the predicted image. Positive and negative patches are those where $\geq 75\%$ and $\leq 25\%$ are white, respectively. We then subtract the mean attention weights of the negative patches from the positive patches. Those with positive differences are therefore more correlated with a positive signal prediction in the cell. For visualization, we depict the log value of the difference (normalized to 1).

Values used to sort candidate NLS sequences are available in the `de_novo_NLS_sequences.csv`.

Predicted sequences are shown in Table S13.

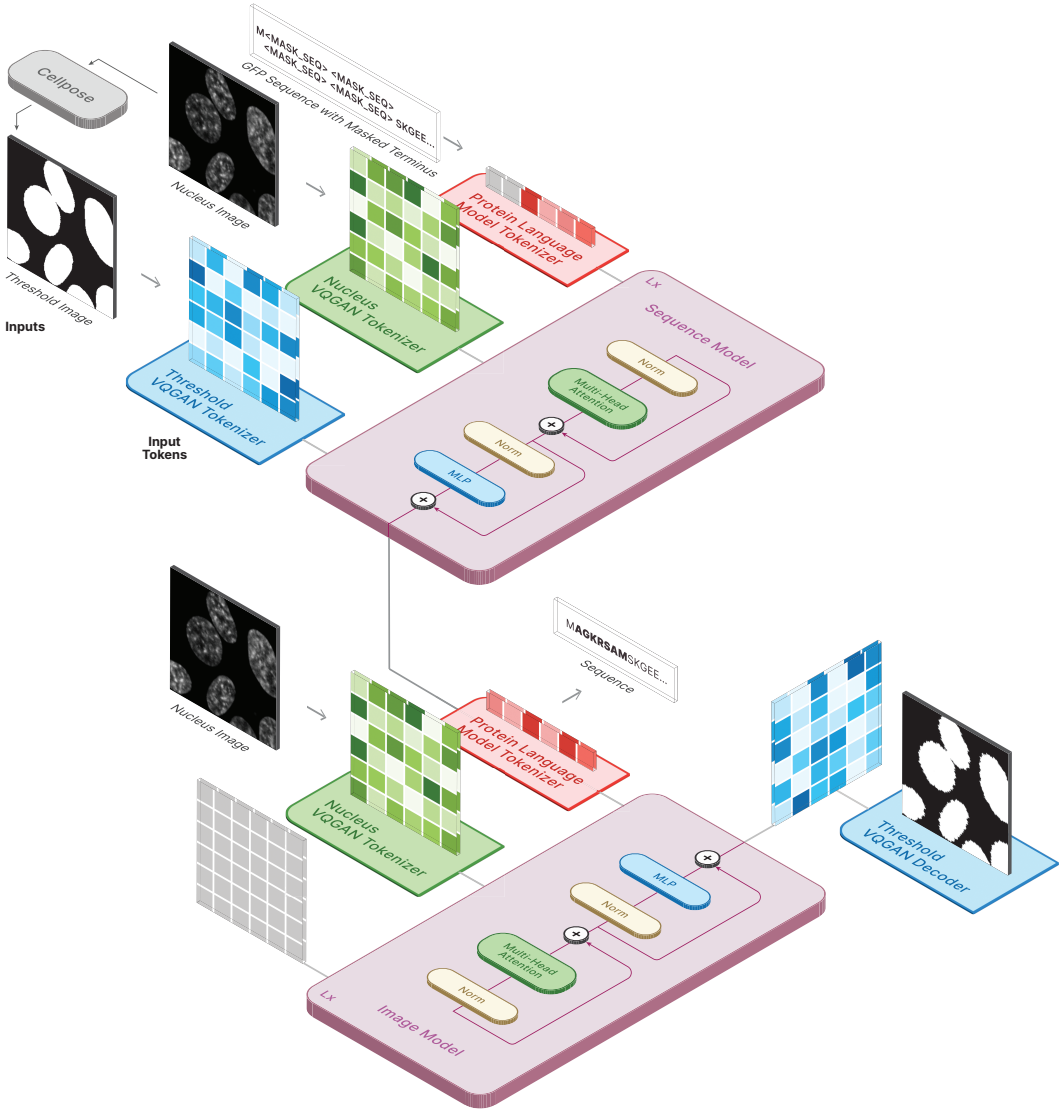

Figure S9: Diagram depicts the pipeline for NLS discovery. In the top half, we predetermine the length of the novel NLS sequence and insert the corresponding number of mask tokens either after the starting Methionine or before the <END> token, depending on the chosen terminus. The threshold image is obtained by passing the nucleus image through Cellpose. In the bottom half, we pass the the GFP with proposed NLS sequence into an image prediction model to ensure predictive consistency of the sequence.

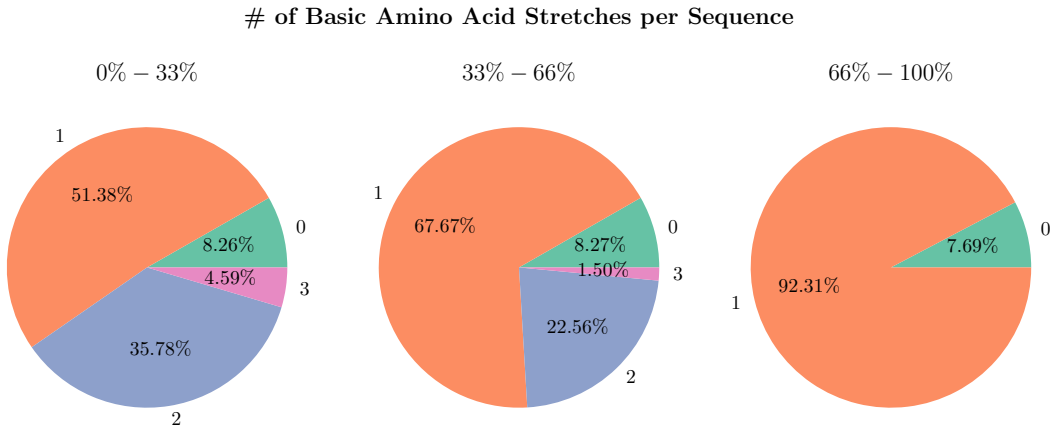

Figure S10: Pie charts showing the maximum # of stretches (numbers outside of circle) of R and K amino acids per proposed NLS sequence. Stretches are calculated based on the number of continuous R and K amino acids with a maximum tolerance of 2 amino acid gap. Only streches with 4 or more amino acids are counted. Proteins are shown binned with respect to Max ID % sequence homology with the NLSdb (0%-33%, 33%-66%, and 66%-100%). The relative proportion of max stretches per bin is shown as a percentage inside the circle.

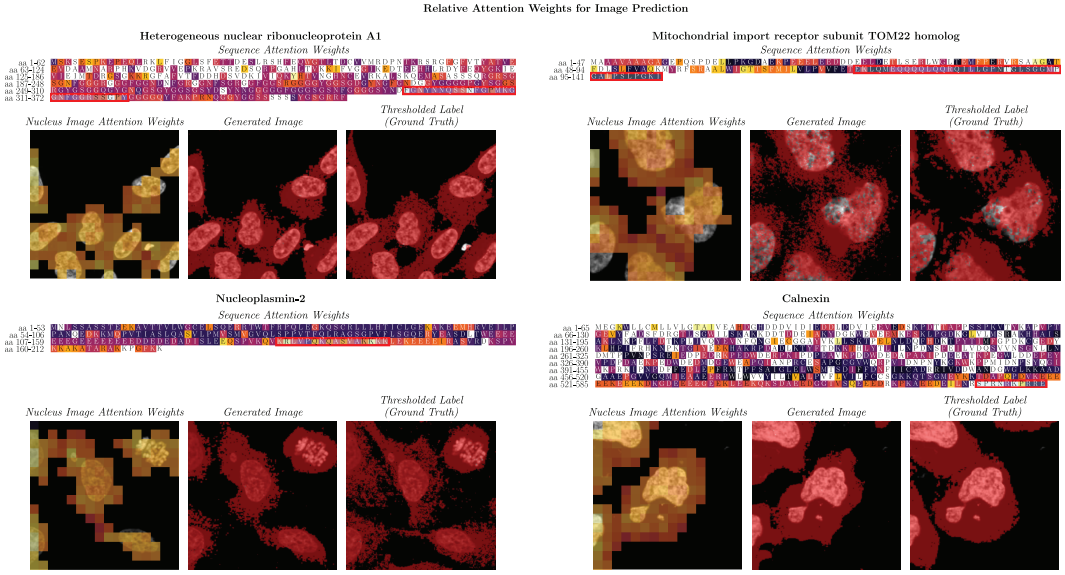

Figure S11: Relative attention weights of predictions from `HPA_480` on HPA images with known localization signals (highlighted in red).
Three proteins with documented localization signals show high attention on those regions: Heterogeneous nuclear riboprotein A1 (top left), which localizes to the nucleus and cytoplasm [18, 19]; Nucleoplasmin-2 (bottom left), which localizes to the nucleus [20]; and Mitochondrial import receptor subunit TOM22 homolog (top right), which localizes to the mitochondria [21]. However, Calnexin (bottom right), which localizes to the endoplasmic reticulum [22], does not show high attention on its localization signal despite the correct prediction. This may be due to the loss of subcellular features in the thresholding process caused by the low resolution of the fluorescence image. We also observe high attention on other amino acids in the sequences that are not known localization signals. These may indicate potential sites of interest for further biological investigation.

Table S13: NLS candidates sorted by nucleus proportion.

Terminus	Sequence	Terminus	Sequence
N	RKRRQR	C	SPTAFPSNVIETIRVKRRMEL
N	NKRPRKKEK	C	EFRAKYRQMGSRKKKKSGQWSA
C	RPKVI	N	KKHKLRSVPDLTELMRMIFLAP
C	VLKRAKKD	N	KLLRFAGKSGMMVLLAPHSGKM
C	RHKKKKIA	C	IFQADKDQKAHPPAKKAPSELMQ
N	HRRKKR	C	KGKVKSIMIPPKSRKSLAKVPLS
C	RSQKRK	N	AAGKSFKPRIKKSRMTRDSSETMA
N	KCKKKN	C	TGNRIFGETPSWERERKRPGGGQQ
N	KGKRFSK	C	NKLQKHSKRQPHKLQAMKLKYPTWE
C	AKRLKGK	C	LVFPNRDASIKKPLQNPPQKRRCMIM
C	SKKAKKNKM	N	LPKRRRLSRRKKVELEPEYGWEEEVVV
C	EEKRPRF	N	TEAPARTAVKKSRAMKGYIARLASSPS
N	MKICIT	C	IEKSKGKEAPKSSPPLKQNQRSRKMVK
N	AVPAKRARIDG	C	FQVRASPKGKPATKNKLRLLKIRRHRV
C	ESHHLPRAKKR	C	LQEGTRTRSQKAQEPKFKKVSGDIPNK
N	GKERSYPPISKR	N	SDPNTAQYPWMPPQATKRAAMAAREAE
C	KLKKRNRQPEDKK	C	HYKKEKRKRSASPILAEEPVPKCARTLR
C	GGKFATGKKKKPKM	C	LDKRKRIKPPKEEQKELMRKMWGPGSSL
N	PSKLLRQ	N	GSKKSRTATDSLESRMAMEDVAMGEESE
C	QRRKGQKFQT	C	EGSGLVPGNSRKRPEPKKPKKRKKVRRK
C	KTCPPKRPVVEW	C	RKKRQAIQAVTMGRIKKKSYEKQWSKFED
C	DKEKKRKNDHEK	C	ASTVPAYSRSKAGKVEPKPKQKKTQRNAP
N	FRFSC	C	SKQQAEINLKAAKPLETTDISLSKKEKKDM
N	LQSSDKK	C	RRAEGLSEPKRHMAEYEQSRRRQRVVRTAT
C	EMEGKKKKIKKM	N	PPTKKQEPQQENNSEDELRRSSSAADPEER
C	LQRKQKMRSH	C	ANFCSGMQAHLSRDFLCL
C	YGEPCIKRSS	C	GNKLARTEMPAVYTSIGSASKSY
C	AQAKRKRIGFH	C	VELRNGKLKPTEESMSFKRMYGS
C	DSSKKPKFTPK	C	EITLSGPPFGGPQVVYRPKLQRVT
C	LKSGPSKSQRKN	C	FGGETQIIENSAKRSHLRPNMHEMI
C	TTKKKKNDSCGAS	C	HKAQPAVIQAISVKRAVEDEPVPMAMT
C	LFGKNRFPKKKKFKM	C	HLTSLKMGGLFVLLPIRSRQKRGSDVG
C	GKKYGHKPRKLKKEK	C	LRDARRSASGLPRQDSEGYVGAPKRIN
N	SAKRGYMLAE	C	LLTGFRLGIGDEKPRRAKHILTSQASK
C	DYPGKGKKRKGKK	C	YVQSIGVEIPGKRGKSSLPSLYQMAEP
N	KRVLHEAPQSAL	C	LKLRLRYNAPIKKLFSRK
C	GPPAKFMLDV	N	PGPSSRYRPLEDGGPAAE
C	SKQACRGKRGSK	C	YPNMPKPRRSKRSVAYTMM
C	DSIPSSRKKRSEM	N	ENEMPTEFHSPKRYQPMNPNS
C	IGPSSSSVEPEFKRT	N	PRNNKKTKMTELGLTQLAEAV
C	IFVQPASDKKRKAMT	N	DSPKRPFVTSVEEPMSMVIMPE
C	SRNRKKRKNRLRRIRKRQFH	C	EIIGNAKRVPEAEGLLHKYQKK
N	PKRRKPMQGGE	C	KASKKVEDQLDAKKPKMEGKAKP
N	KRALMAEPVVE	C	TQEKAQKKADLRGQPQRKRSKEM
C	KKEKVSKRKQRRRF	C	KPQEVLKEIECTQKPTKKKVLDG
C	VEGKGMKRSVRAV	C	IATATHRKRGIKHPHRRRSRPLFG
C	RQRPAYNAVDI	N	QSNYKRQKVPPPENSEMRVAMGSEL
C	TYKKLPTDKKQQQILKR	C	FSKKPEPTGKRPKKSSRSKFRCHRN
C	FALKQDHKKAK	N	KRKTNQIPSKREGDQTNMADTKRQKL
N	GNHKRYKMKERMGLF	C	FRTKPPKGKNRMSETGSFAMAVKANA
C	KKWKQRIKRILPLI	C	TKEPKKPHKKTKMRLRRLNGNSESMSC
N	ELGERPGSRKRTGRE	C	DWFTYAQNQAVSNAIEEHHSMLKKHKI
N	SLTKAFSQMQRSQKK	C	DLRNRRLHLSKVEIVWYGALSKQPRTN
N	LKLHSKLLEKKNKRMM	C	RKRRRGLDRPGYNSSTSHGDDPPTSGW
C	VTLDQTKKSKTRRKHIFR	C	HALRKGRIELVYKQTKRSAAITSRYTLE
N	DASEMLKGKLKKMKSEGLT	C	KRKAAEDTTEVEMSPGGDEEEKHASPSS
C	GNRKAKRKDGTLDRNHRLEN	C	DKDNLCLKKRERLEDMGYLPKKRASAMRM
C	LDANGPFKDMVKNKRAKRQC	C	NIKLDDPIPTDRTGEILMDARKSKIRPMM
C	RDFKEPKPKRRRIRRASGAP	C	QSLDPKDDDSAKRPALPHPAKAIKKSRLH
C	DRAVLPPPYKHQKRKEATKKKM	C	NPTLHAPIHFGKMRNLTPPPPPTKKKMKP
C	AYKLRGVESASAPHSPIKRKEM	N	PRPSLAKRPRFVACKQLMLPDDPVSLHYK
N	PTPPSKRQPELSLEFAKQAAREA	N	PPKQRRRHKTDESLFLGRPDTPSVEWKRKQ
C	KKKRPGRARRRRRKKKQGELKIQH	C	SKSPMLAGGGEPHDPSGTESEPVSMRTHTM
C	KFRGGKKRKRRTDKKTQSVTRKRRK	N	AEEELTVAVTTASEPAWAGMSSITEIAAKR
C	VKYEPGFSRQQGRI	C	KKDAAAPGLVTGDEKRTAM
C	RQKLSYALVEGMVD	C	VPPGYRDKDVKRAKPLSPSYVA
C	SRAKRKAEPIVWVLA	C	RKSRKILCPYMRFYFEHATVGAW
C	APIFVESPQSSGQNKRE	C	KKENTPVQLVPPSKKAARTSLISK
C	KKRGRWGRIRPSYVKDKCL	C	SVSKRSRDLVPWSEEGFFQQAKQIQ
C	LLSDSSSLQHALEPKKIQI	C	NVRPAIKKQIPLYDLQRQPEKMRKLINM
C	NTTKPKRKQNKTIT	C	DFKKKRKKWLLARRMQAC
N	PPSRGKKLTDNRRSKSPSPLPE	N	ELAREQEMSPAKRHMTWGTL
N	DPGPAKKARTMTQS	C	PHKITEDLTQERRKRGKGGH
C	NENPTVKQECKK	C	IGAAKKLHQPVGERASKKAMM
C	KEYIKYQKKKLMM	N	SSTEPPADPSAPRSKIPRLATE
C	MIKPAKRSKTEKPQN	C	LVLEKSASSVMEAPSKILKQKM
C	AKKFESLAMKFQRLN	C	AASPLPLEPPANLGDRKKRKEAIK

C	RPTVLPKPGSRQAKKSY
C	KVEDIEPNTKKFSGKQS
C	KRSKGKMWMKNLFPEKL
C	RKKKKKSRTEREPIRKRK
N	STKRCEVERSENLDAGEM
C	EPVGSTKFRKRQKIRGISN
C	KYRSKKAFREMRTKVGGGM
C	KVSDKASEQHARRKKRQSS
N	GKHTCSNKGKRKRKLIHFKSRM
C	DRKKDITGHGPEKKKLRKEQQK
C	NVDNENIDKKKKFKSVTKGKHD
C	ATEGKEPVGPGSSKGRRRRRRRP
C	QGIEVDSSIKGFSHKKKKRKMKM
C	RRKNKLRPARRRRLYPSKRRRRLRPN
C	ERAATAAASTSTKEASPPASKKSYKFEF
C	DKKGRKPGRSTGVI
C	KRAARRSRVVAPIRSI
C	HSSGSPLEKLGRKNNRNRAS
C	RTRVDGAAAASE
N	AMAGQTKRRPQRKA
N	SGDGPFHQSKGKRKH
C	WNCKRLKEKKSEHPAA
N	SGPPAKVQKRAPESDCR
C	RHPPAEETPKAAKRKPTI
C	DKETSKDIGRGGRGKRKLDL
C	KKKKKQRKKRKRDQGRLRKW
C	LSFERGKMKRLHKKKRKIKL
C	KGKTYYKRVRERMPKKRPLT
C	KKREKRKQKEAKHKRRIKSMLE
C	SMPKELNSLVPKKRRQGPVRQDTQ
C	PQSKRDGKQKDSDN
C	RGEAKKESENAKRHQ
C	KEKQNITKKAKRKTHK
C	DRKSNPFVFLKPKTEEM
C	KRKDKQIAVKKYPRTKS
C	KLLKTTKITKDAKYPRKH
C	ARYSKSKKKFYNSKLMPH
C	RGGKKKKGARAPVFGASLD
C	VDVAFVHKSPGSRKQRGRF
C	RTTKKRQTRPPAPRDRRNSL
C	SKLEEKKWALLSSQKHTRQG
N	NKKKNKTCAAAPAAAAPTVM
N	SKKKKYPGILRVPVGQLPLAEMKSA
N	PKKKRKAPAVWQAAEPAPSSMPPVE
N	PFLLVSQLG
N	LLATAGIYHLL
C	HSSKHLARVL
C	RVCRKGNMFIDSSKERS
N	MMMMMMMMMKMMMMLCQTLTGQRKRG
C	FLRINAVHRAKGPKKIKSLPA
C	HPKKKRATGWSPKKQASRKRPKWNAI
C	KGESSGKKQTLKKVCLGHEKRTFSKA
C	ASSKCDHNERDRSSRDKRKTSKKKGNK
C	YFSISRTISKTRKARPRGWEGSKKSRMM
C	STISSVATRRSKKEQRMPAAPSNNLPKKI
N	PRRRREADVETRDAAMGGEPKVLQVLHLGN
C	IRYMNIPQRGIPKLPRSE
C	SFTHQDNMPSKRFNGRGRMQH
C	KQRAATLKQTSEESKKPRPIDLH
C	AAPSALSREEPGLWGSMAKRTVLA
C	TSKDQPPHKLMQRAV
N	TMMAMQLARRMGPRFMRSSF
C	KSKFKRQKYAGDHGLKEGDI
C	VPAEREKNRRKRQTHLGYSMGL
C	DVMPNKKLCIVLPPKSLSDAPMQ
C	PLETDHMHRTWSTKIRMCVLMIT
C	KLKRRGIIITGETLNESGLKKLA
N	AARKRGQAKLLERRLEWFWMMIGDML
C	RQSQSISAKWKRESAASQSGEQAEMNM
C	QVRKRYYVRLTSEKPKIPKYQKWLYWM
C	LCMDIVIEYTDARIRKKTAKFLKEINE
C	IYPGKEPPIKLNKSLKSKRESHSADMSF
N	EVSKAQRKQKPAKLPPSTTIQIASVDYE
N	KGGRKEVEVQQRESAPLPALPSEAYEEAVE
C	NMLSPSEPSYVGSTKYGKSIR
C	VHSPWMGVSSTEGLLFLPVKILKQV
C	YAQEPELQSKFKAQRLTDPYFYGPH
C	SRGLAWLMPTVLLCPHKPFRLRVDS
N	RIGSIWEFVRRKEQFWLRVTAMA
C	KLLIEPYAKAKKNWISMLCSAAMGSFL
C	KSRNKTPPKKGLCVVTSSLKKTVTMTKS
C	SIFGDGKLKDARRKVPHKRRLRILFLSYC
C	GSGLRRKSTKTLQQTSDMAEGKS
C	TLIPFHALKNIFAVVALQALRVVG
C	AALIGSASPLALLRHGVQVLSPDSYW
C	KYKGEQTIVKQEHLGDGVVARMPT
C	RKEMFVRPPTHTHTVTMILRKKLKLSAS
C	SNRHAIMSRPEYNKHEDDNKMQKYIVWM
N	AGASLVMDTAGIGGSVGMRIQTKRHKVD
N	KRFMPMMSQNTIHNNPQYINARPSRFPLY
N	ATAHPTSNASWEKESAHAPVKKVHRMKEP
C	REHKPAQQQAKGKEPKVPPPTGERTMGYQ
C	AAKKSRTLPETKSGGMKTVRLLEGPMDF
N	AAATNPTRAMITLKENRKGHMMGKNKKA
C	VDKKLPPKECMKKMIKMAISKLVAKPTK
C	YTSVTNFGFKAHDLDFGKFKQEPDLDYD
N	ISFSKILMLPLMSLSTAPAMKVQHED
N	AMMAVAMMTMVAMGQFAGDTLKKRNRGE
N	LAIGAVEPAMAQEPMIETTMVFQVPERS
C	DGTKLLEGQFTKQSCAATILFPSHD
N	AMAGLAYGQENVPPKNGQGQT

References

[1] Andreas Digre and Cecilia Lindskog. The Human Protein AtlasSpatial localization of the human proteome in health and disease. *Protein Science*, 30(1):218–233, 2021. ISSN 1469-896X. doi: 10.1002/pro.3987. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3987. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.3987.

[2] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, January 2015. ISSN 0305-1048. doi: 10.1093/nar/gku989. URL https://doi.org/10.1093/nar/gku989.

[3] Nathan H. Cho, Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y. S. Kim, Edna M. Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel M. Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin B. Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loic A. Royer, Matthias Mann, and Manuel D. Leonetti. OpenCell: proteome-scale endogenous tagging enables the cartography of human cellular organization. Technical report, March 2021. URL https://www.biorxiv.org/content/10.1101/2021.03.29.437450v1. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[4] Middi Venkata Sai Rishita, Middi Appala Raju, and Tanvir Ahmed Harris. Machine translation using natural language processing. *MATEC Web of Conferences*, 277:02004, 2019. ISSN 2261-236X. doi: 10.1051/matecconf/201927702004. URL https://www.matec-conferences.org/articles/matecconf/abs/2019/26/matecconf_jcmme2018_02004/matecconf_jcmme2018_02004.html. Publisher: EDP Sciences.

[5] Nicholas C. Bauer, Paul W. Doetsch, and Anita H. Corbett. Mechanisms Regulating Protein Localization. *Traffic*, 16(10):1039–1061, 2015. ISSN 1600-0854. doi: 10.1111/tra.12310. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tra.12310. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tra.12310.

[6] Juane Lu, Tao Wu, Biao Zhang, Suke Liu, Wenjun Song, Jianjun Qiao, and Haihua Ruan. Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell Communication and Signaling*, 19(1):60, May 2021. ISSN 1478-811X. doi: 10.1186/s12964-021-00741-y. URL https://doi.org/10.1186/s12964-021-00741-y.

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked Generative Image Transformer, February 2022. URL http://arxiv.org/abs/2202.04200. arXiv:2202.04200 [cs].

[8] Huiwen Chang, Han Zhang, Jarred Barber, A. J. Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-To-Image Generation via Masked Generative Transformers, January 2023. URL http://arxiv.org/abs/2301.00704. arXiv:2301.00704 [cs].

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. October 2017. URL https://openreview.net/forum?id=BJJsrmfCZ.

[10] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv:2104.09864 [cs]*, October 2021. URL http://arxiv.org/abs/2104.09864. arXiv: 2104.09864.

[11] Emaad Khwaja, Yun S. Song, and Bo Huang. CELL-E: Biological Zero-Shot Text-to-Image Synthesis for Protein Localization Prediction, May 2022. URL https://www.biorxiv.org/content/10.1101/2022.05.27.493774v1. Pages: 2022.05.27.493774 Section: New Results.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, December 2014. URL https://arxiv.org/abs/1412.6980v9.

[13] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, November 2017. URL https://arxiv.org/abs/1711.05101v3.

[14] Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature Methods*, 19(12):1634–1641, December 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01663-4. URL https://www.nature.com/articles/s41592-022-01663-4. Number: 12 Publisher: Nature Publishing Group.

[15] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8, January 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL https://www.nature.com/articles/s41587-022-01618-2. Publisher: Nature Publishing Group.

[16] Michael Bernhofer, Tatyana Goldberg, Silvana Wolf, Mohamed Ahmed, Julian Zaugg, Mikael Boden, and Burkhard Rost. NLSdb-major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Research*, 46(D1):D503–D508, January 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx1021.

[17] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers, May 2020. URL http://arxiv.org/abs/2005.00928. arXiv:2005.00928 [cs].

[18] Lars Jønson, Jonas Vikesaa, Anders Krogh, Lars K. Nielsen, Thomas vO Hansen, Rehannah Borup, Anders H. Johnsen, Jan Christiansen, and Finn C. Nielsen. Molecular composition of IMP1 ribonucleoprotein granules. *Molecular & cellular proteomics: MCP*, 6(5):798–811, May 2007. ISSN 1535-9476. doi: 10.1074/mcp.M600346-MCP200.

[19] Qing Liu, Shi Shu, Rong Rong Wang, Fang Liu, Bo Cui, Xia Nan Guo, Chao Xia Lu, Xiao Guang Li, Ming Sheng Liu, Bin Peng, Li-ying Cui, and Xue Zhang. Whole-exome sequencing identifies a missense mutation in hnRNPA1 in a family with flail arm ALS. *Neurology*, 87(17):1763, October 2016. doi: 10.1212/WNL.0000000000003256. URL http://n.neurology.org/content/87/17/1763.abstract.

[20] J. P. Makkerh, C. Dingwall, and R. A. Laskey. Comparative mutagenesis of nuclear localization signals reveals the importance of neutral and acidic amino acids. *Current biology: CB*, 6(8):1025–1027, August 1996. ISSN 0960-9822. doi: 10.1016/s0960-9822(02)00648-6.

[21] M. Yano, N. Hoogenraad, K. Terada, and M. Mori. Identification and functional analysis of human Tom22 for protein import into mitochondria. *Molecular and Cellular Biology*, 20(19):7205–7213, October 2000. ISSN 0270-7306. doi: 10.1128/MCB.20.19.7205-7213.2000.

[22] Asvin Kk Lakkaraju, Laurence Abrami, Thomas Lemmin, Sanja Blaskovic, Béatrice Kunz, Akio Kihara, Matteo Dal Peraro, and Françoise Gisou van der Goot. Palmitoylated calnexin is a key component of the ribosome-translocon complex. *The EMBO journal*, 31(7):1823–1835, April 2012. ISSN 1460-2075 0261-4189. doi: 10.1038/emboj.2012.15. Place: England.