CELL-E 2: Translating Proteins to Pictures and Back with a Bidirectional Text-to-Image Transformer

Emaad Khwaja[*‡] emaad@berkeley.edu

Yun S. Song[†‡] yss@berkeley.edu

Aaron Agarunov[§] agarunov.aaron@gmail.com

Bo Huang [¶‖**] bo.huang@ucsf.edu

Abstract

We present CELL-E 2, a novel bidirectional transformer that can generate images depicting protein subcellular localization from the amino acid sequences (and *vice versa*). Protein localization is a challenging problem that requires integrating sequence and image information, which most existing methods ignore. CELL-E 2 extends the work of CELL-E, not only capturing the spatial complexity of protein localization and produce probability estimates of localization atop a nucleus image, but also being able to generate sequences from images, enabling *de novo* protein design. We train and finetune CELL-E 2 on two large-scale datasets of human proteins. We also demonstrate how to use CELL-E 2 to create hundreds of novel nuclear localization signals (NLS). Results and interactive demos are featured at https://bohuanglab.github.io/CELL-E_2/.

1 Introduction

Subcelllular protein localization is a vital aspect of molecular biology as it helps in understanding the functioning of cells and organisms [1]. The correct localization of a protein is critical for its proper functioning, and mislocalization can lead to various diseases [2]. Protein localization prediction models have typically relied on protein sequence data [3, 4] or fluorescent microscopy images [5, 6] as input to predict which subcelleular organelles a protein would localize to, designated as discrete class labels [7, 8]. The CELL-E model was markedly different in that it utilized an autoregressive text-to-image framework to predict subcellular localization as an images [9], thereby overcoming bias from discrete class labels derived from manual annotation [10]. Furthermore, CELL-E was capable of producing a 2D probability density function as an image based on localization data seen throughout the dataset, yielding more a far more interpretable output for the end user.

Although novel, CELL-E was inherently restricted by the following limitations:

Autoregressive Generation. Alongside other autoregressive models [11–14], CELL-E was limited by slow generation times and unidirectionality. When provided with input, CELL-E required a separate step for each image patch (256 for the output image composed of tokens of size 16×16).

[*]UC Berkeley - UCSF Joint Bioengineering Graduate Program
[†]Department of Statistics, UC Berkeley, CA 94720
[‡]Computer Science Division, UC Berkeley, CA 94720
[§]Department of Pathology, Memorial Sloan Kettering Cancer Center, 10065
[¶]Department of Pharmaceutical Chemistry, UCSF, San Francisco, CA 94143
[‖]Department of Biochemistry and Biophysics, UCSF, San Francisco, CA 94143
[**]Chan Zuckerberg Biohub - San Francisco, San Francisco, CA 94158

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

This slow image generation severely limits the ability of CELL-E to perform a high-throughput mutagenesis screening.

Unidirectional Prediction. The unidirectional nature of CELL-E allowed for predictions to be made in response to an amino acid sequence, however it may be of interest to biologists to make predictions of sequence given a localization pattern. Such capability would be advantageous for those working in a protein engineering domain [15, 16]. One could imagine a researcher wanting to know the optimal localization sequence to append to a protein on either the N or C terminus [17] while maintaining essential function within an active site region, as well as reducing the chance of off-target trafficking.

Limited Dataset. CELL-E utilized the OpenCell dataset [18], a relatively small dataset. Vision transformers often require large amounts of data to make robust predictions [19], however a small dataset was utilized in the original model. This led to a degree of overfitting and prediction bias based on the limited diversity in localization patterns of the original dataset.

Present Work. As in CELL-E, our method CELL-E 2 is able to generate accurate protein localization image prediction as illustrated in Fig. 1, but it differs from CELL-E by employing a non-autoregressive (NAR) paradigm which improves the speed of generation. Similar to CELL-E, we retrieve embeddings from a pre-trained protein language model and concatenate these with learned embeddings corresponding to image patch indices coming from a nucleus (a subcellular organelle containing DNA [20]) image and protein threshold image encoders (Fig. 2). We then apply masking to both the amino acid sequence as well as the threshold image in an unsupervised fashion, and reconstructed tokens are predicted in parallel, allowing for generation in fewer steps. This also allows for bidirectional prediction, (sequence to protein threshold image or protein threshold image to sequence). Additionally, to improve the predictive performance we utilize a larger corpus of data, the Human Protein Atlas (HPA) [21] in pre-training to expose the model to a higher degree of localization diversity, and finetune on the OpenCell dataset [18], which better represents natural protein localization because it is acquired from live instead of fixed cells. We explore multiple strategies towards finetuning which serves to generally inform task-specific refinement text-to-image models in Section 5.3.

2 Related Work

2.1 Protein Language Models

Embeddings from unsupervised protein language models can be used to predict and analyze the properties of proteins, such as their structure, function, and interactions [22]. By exploring the hidden patterns and relationships within these sequences, protein language models can help to advance our understanding of the complex world of proteins and their roles in various biological processes. Masked language modelling has been particularly successful. Ankh [23], ProtT5 [24], ProGen [25], ESM-2 [26], and OmegaFold [27] are examples of recent models which use masked langauge approaches. Embeddings from ESM-2 and Omegafold in particular have been used for structural prediction, indicating hierarchies of information beyond the primary sequence contained in the embeddings [28].

2.2 Protein Localization Prediction

Protein localization prediction via machine learning is an emerging field that uses computational algorithms and statistical models to predict the subcellular spatial distribution of proteins [29]. This is an essential task in biology, as the subcellular localization of a protein plays a crucial role in determining its function and interactions with other proteins [30, 31] The prediction of protein localization is performed by analyzing protein sequences, amino acid composition, and other features that can provide clues about their subcellular location. Machine learning algorithms are trained on large datasets of labeled proteins to recognize patterns and make predictions about the subcellular location of a protein. This field has the potential to improve our understanding of cellular processes, drug discovery, and disease diagnosis.

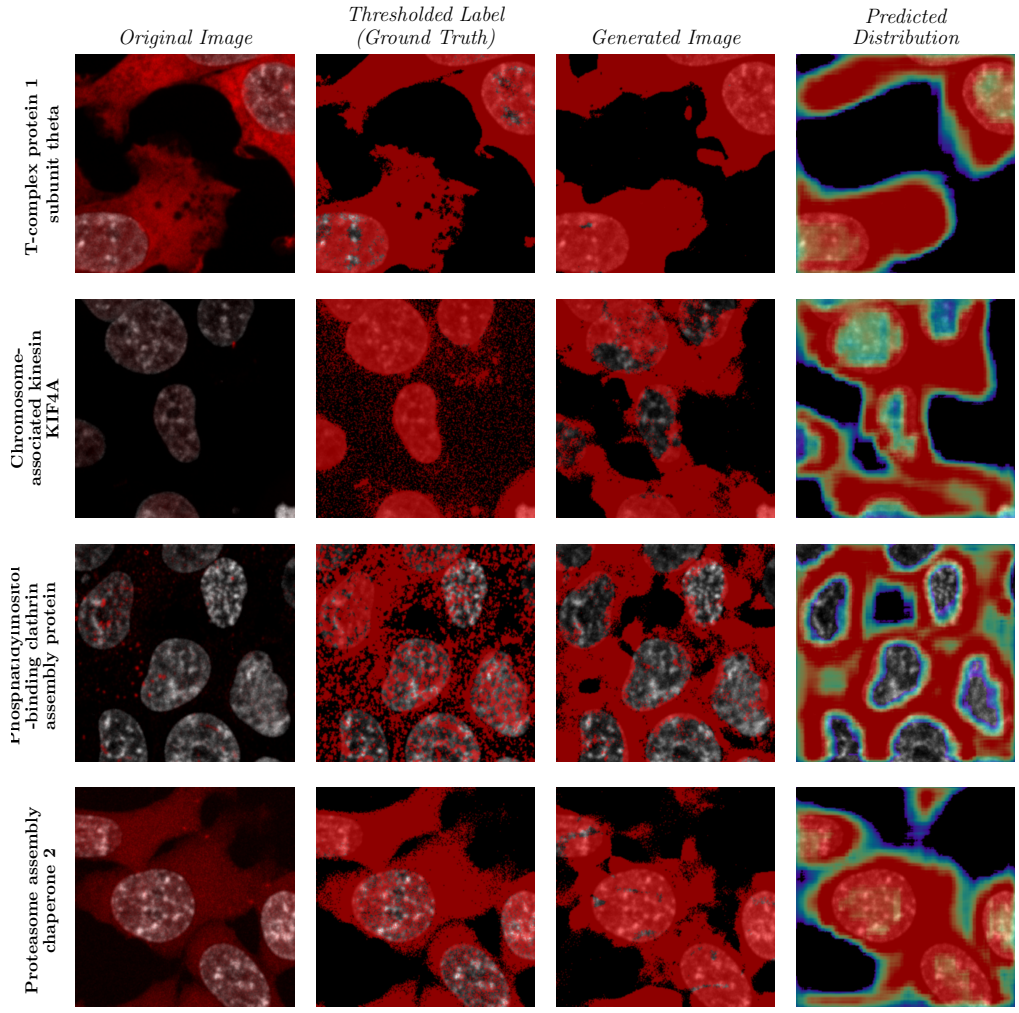

Figure 1: Localization predictions from CELL-E 2 (`HPA Finetuned (Finetuned HPA VQGAN)_480`) on randomly chosen validation set proteins from the OpenCell dataset. All images feature the Hoescht-stained nucleus image as a base. The "Original Image" column shows the fluorescently labelled protein from the dataset. The "Thresholded Label" shows pixels greater than the median value. This serves as the ground truth for the model during training. "Generated Image" is the image specifically predicted by CELL-E 2 and is compared against the thresholded ground truth image. "Predicted Distribution" is the latent space interpolation of the binary threshold image tokens which uses which utilizes the output logits of CELL-E 2. See Fig. S1 for colorbars corresponding to all plots in this work.

Recently, attention-based methods have demonstrated the ability to predict localization from a sequence [32], enabling the use of long context information when compared to convolutional-based counterparts [33–35]. These methods, however, predict localization as discrete classes rather than as an image. CELL-E, on the contrary, does not utilize existing annotation and provides a heatmap of the expected spatial distribution on a per-pixel basis [9]. This approach enables learning at scale by eliminating the bottleneck of manual annotation while also circumventing label bias.

2.3 Text-to-Image Synthesis

Recent gains in the text-to-image field have have largely been made by autoregressive models [11, 13], which correlate text embeddings with image patch embeddings, as well as diffusion models, [14, 36–39], which condition on sentence embeddings to gradually synthesize images from random noise.

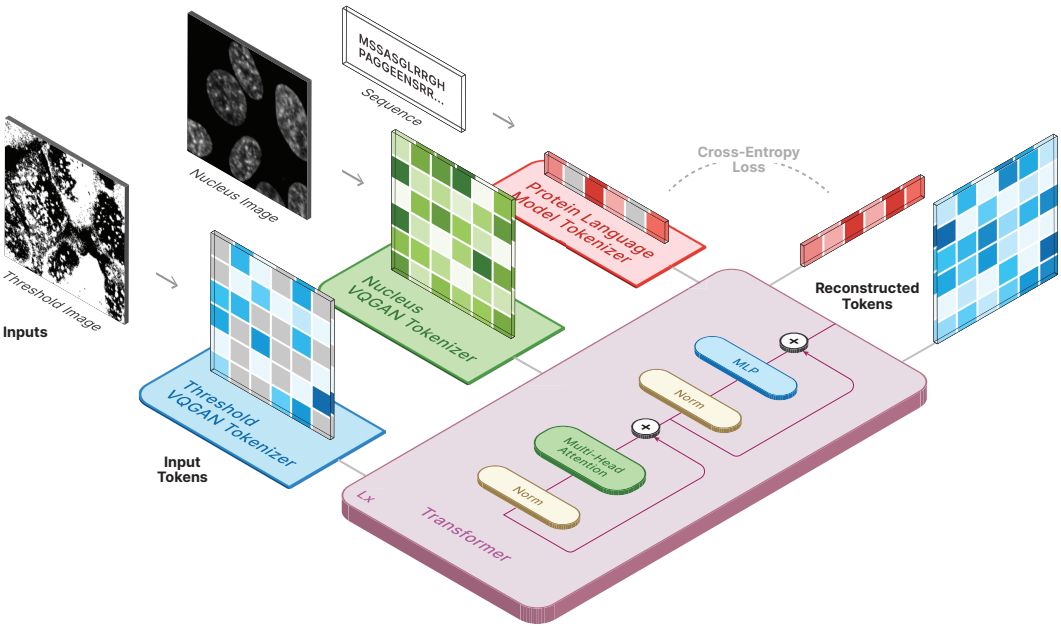

Figure 2: Depiction of training paradigm for CELL-E 2. Gray squares indicate masked tokens. Loss is only calculated on masked tokens in the sequence and protein threshold image.

Other works implement non-autoregressive models (NAR), which take advantage of a masked reconstruction procedure, similar to BERT [40], where the model is tasked with predicted randomly masked portions of an input image. These types of models are particulalry advantageous because they enable parallel decoding, allowing images to be synthesized in relatively view steps when compared to autoregressive models. Furthermore, NAR models are not bound to a particular direction of synthesis like autoregressive models, which only perform next-token prediction. CogView2 [41] utilizes a modified transformer architecture where attention on masked tokens is eliminated. MUSE [42] builds on MaskGIT [43] by concatenating a pre-trained text embedding to a token masked representation of a corresponding image. It uses a vanilla transformer architecture [44] and yielded state-of-the-art image synthesis performance in terms of FID and human evaluation.

3 Datasets

We pretrained our model on protein images from the Human Protein Atlas (HPA) [45], which covers various cell types and imaging conditions using immunofluorescence microscopy1. We then finetuned on the OpenCell dataset [18], which has a consistent modality using live-cell confocal microscopy of endogenously tagged proteins. To ensure generalization to new data, we followed the homology partitioning method of Almagro Armenteros et al. [35]. We used PSI-CD-HIT [46] to cluster HPA proteins at ($\geq 50\%$) sequence similarity and randomly split the clusters into 80/20 train/validation sets. We applied the same clustering and splitting to the OpenCell proteins, matching the train/validation labels from HPA. For proteins present in OpenCell but not HPA ($= 176$), we assigned the protein based on the other labels in the cluster. Any remaining unassigned proteins ($= 1$) were assigned to the training set. See Appendix A for more details about the datasets and preprocessing.

4 Methods

CELL-E 2 (Fig. 2) is a masked NAR transformer model, which upgrades the capabilities of CELL-E, an autoregressive *decoder-only* model [47]. Due to the NAR nature of the model, CELL-E 2 is capable of both image generation (sequence to image), as well as sequence prediction (image to sequence).

4.1 Amino Acid Sequence Embeddings

Proteins are biological molecules which are comprised of individual amino acids. CELL-E 2 utilizes embeddings from ESM-2 [26], where amino acid molecules are denoted with individual letter codes (e.g. A for alanine) [48]. We opt to use frozen amino acid embeddings for the prediction task, which has been demonstrated to yield superior reconstruction performance in text-to-image models [9, 37, 42]. The embeddings obtained from a protein language model contain valuable information about amino acid residues, biochemical interactions, structural features, positional arrangements, as well as other characteristics like size and complexity [22]. We train models of varying size based on the released ESM-2 checkpoints (See Section 5). The output of the final embedding layer per respective model is used as the amino acid sequence embedding.

4.2 Image Tokenization

Just as in Khwaja et al. [9], we utilize a nucleus image, which serves as a spatial reference with which a binarized protein threshold image is associated. We chose this in order to parallel the type of images which are typicall acquired in a wet lab scenario.

We also utilize VQGAN autoencoders [49] trained on both the HPA and OpenCell datasets, respectively. VQGAN surpasses other quantized autoencoders by incorporating a learned discriminator derived from GAN architectures [50]. Specifically, the Nucleus Image Encoder employs VQGAN to represent 256×256 nucleus reference images as 16×16 image patches, with a codebook size of $(n = 512)$ image patches. To enable transfer learning, we explore finetuning strategies these VQGANs in Section 4.5.

The protein threshold image encoder acquires a compressed representation of a discrete probability density function (PDF) that maps per-pixel protein positions, presented as an image. We binarize the image based on the median pixel value of the image (see Appendix A.4). We utilize a VQGAN architecture identical to the Nucleus VQGAN to estimate the entire set of binarized image patches to denote local protein distributions. These VQGANs are trained until convergence, and the discrete codebook indices are used for the CELL-E 2 transformer. Hyperparameters (Table S1, Table S2, Table S3) and training details can be found in Appendix B.2.

4.3 Input Masking Strategy

We adopt a cosine-scheduling technique for masking image tokens, which has been used by other works. The probability of an image patch being masked is determined by a cosine function, favoring high masking rates with an expected masking rate of 64% [42, 43]. This technique provides various levels of masking during the training process, exposing the model to spatial context for masked language tokens.

Of similar interest as image prediction, sequence in-filling with respect to a localization pattern has potential to significantly augemnt a biologist's workflow. Typically, protein localization sequences are found through sequence similarity searches with proteins that have known localization in particular organelles [51–53] or via experimentation [54, 55]. CELL-E 2's bidirectionality enables the model to make predictions for image with respect to amino acid sequences, as well as sequence predictions conditioned on images. This enables an entirely new paradigm in protein engineering which relies on image information. To achieve this, we also mask the language tokens along with the protein threshold image tokens. We experimented with using the same cosine function for image masking but found it led to numerical instability and vanishing gradients. Therefore, we linearly scaled the cosine function to ensure the maximum masking rate matched 15% masking rate used to train ESM-2.

4.4 Base Transformer

The base transformer is an NAR model in which the embedding dimension is set to the embedding size of the pretrained language model used. We utilized two types of masking tokens. For masking the amino acid sequence, we leveraged the mask token which already exists within the ESM-2 dictionary, designated as `<MASK_SEQ>`. The VQGAN does not contain a masking token within its codebook, so to represent it, we add an additional entry in the image token embedding space of CELL-E 2 (with $n+1$: $(512+1 = 513)$, where n is the number of tokens in the VQGAN codebook),

and designate the final token as <MASK_IM>. We additionally create an embedding space of length 1 for the <SEP> token which is appended to the end of the amino acid sequence. Training details can be found in Appendix B.2.

We sample from this transformer by strategically masking positions in the image or sequence (see Appendix B.1). The logit values for the image prediction are used as weights for the threshold image patches to produce a predicted distribution (Fig. 1, Fig. S5).

4.5 Finetuning

We sought to leverage useful information from both HPA and OpenCell. HPA contains many proteins (17,268), but is potentially subject to inaccuracies, fundamentally because of the immunohistochemistry used for staining requires several rounds of fixation and washing [21]. This means the proteins are not observed in a live cell; are subject to signal loss, artifacts, and/or relocalization events; and therefore may not necessarily represent the true nature of protein expression and distribution within a cell [56]. The OpenCell dataset, while comparatively smaller, overcomes these issues by using a split-fluorescent protein fusion system allows for tagging endogenous genomic proteins, maintaining local genomic context, and the preservation of native expression regulation for live cell imaging [18, 57]. We therefore utilize the HPA dataset for pretraining, and then finetuned on OpenCell.

Optimally finetuning withinin the text-to-image domain remains an open question. The use of multiple models makes it difficult to pin down the correct strategy. Contemporary efforts utilize pretrained checkpoints to fine-tune on domain specific data [58–60]. Chambon et al. [61] reported improved synthesized image fidelity when fine-tuning the U-net of a text-to-image diffusion model, but similar fine-tuning strategies have not been explored for patch-based methods. We report our findings in Section 5.3.

5 Results

Similar to CELL-E, we cast the embedding spaces for the image tokens at the same size as the ones used by the pre-trained language model. The size of the embedding vectors ("Hidden Size") for each model was chosen based on the publicly available ESM-2 checkpoints. For instance, a CELL-E 2 model with hidden size $= 480$ uses `esm2_t12_35M_UR50D`, which corresponds to a 35M parameter model with 12 attention layers. Khwaja et al. [9] demonstrated a positive relationship between the number of attention layers (designated "Depth") in the base transformer and the image prediction performance. The maximum depth was set based on our available GPU memory capacity. We refer to models using the name format "`Training Set_Hidden Size`".

5.1 Protein Localization Image Prediction Accuracy

To predict the protein localization image, we provide CELL-E 2 with the protein sequence and nucleus image, and fill the protein image token positions with <MASK_IM> tokens (Fig. S3).

We evaluated the models on several image metrics (see Appendix C.1) that measure the quality and diversity of the generated protein images (Table 1). Additionally, we assessed the model's generalization capabilities by testing them on the other dataset (HPA-trained model on OpenCell and *vice versa*) (Table S4). We report the results for each model on its respective dataset. We observed a significant positive effect of depth on performance across all metrics and datasets. The models with hidden sizes of 480 and 640 achieved the highest scores, with no significant difference between them. However, on the HPA dataset, `HPA_640` surpassed the `HPA_480` model in more categories. On the OpenCell dataset, `OpenCell_480` performed better than the `OpenCell_640`.

We conducted a visual examination of the generated protein images as depicted in Figures Fig. S6 and Fig. S7. Among the models, OpenCell demonstrated a higher visual resemblance and consistency with its respective ground truth labels, although they exhibited low entropy in the predicted distribution. This indicates that while these models accurately identified the correct tokens with high probability, they struggled to account for the uncertainty and variety inherent in other valid choices, possibly due to a tendency for rapid overfitting which hindered their generalizability.

Table 1: Validation Set Image Prediction Accuracy. MAPE: mean absolute percentage error, MAE: mean absolute error, SSIM: structural similarity index measure, FID: Fréchet inception distance, IS: inception score.

Dataset	Hidden Size	Depth	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
HPA	480	68	**0.0257 ± 0.0250**	0.3340 ± 0.0788	0.2846 ± 0.0985	0.2633 ± 0.1781	12.0332	2.2900 ± 0.0410
	640	55	0.0294 ± 0.0278	**0.3283 ± 0.0805**	**0.2842 ± 0.0991**	**0.2826 ± 0.1827**	21.7942	2.2618 ± 0.0364
	1280	25	0.0370 ± 0.0360	0.3622 ± 0.0799	0.2967 ± 0.0985	0.2645 ± 0.1857	**1.5161**	**2.5440 ± 0.0490**
	2560	5	0.0818 ± 0.0794	0.3516 ± 0.0792	0.3104 ± 0.0904	0.2558 ± 0.1619	23.7977	2.1578 ± 0.0290
OpenCell	480	68	0.0161 ± 0.0148	**0.4953 ± 0.0064**	**0.3620 ± 0.1168**	**0.1220 ± 0.1188**	**1.5844**	**2.6069 ± 0.1175**
	640	55	**0.0159 ± 0.0136**	0.4995 ± 0.0006	0.3785 ± 0.1008	0.1011 ± 0.1012	2.6966	2.0974 ± 0.0981
	1280	25	0.0272 ± 0.0223	0.4996 ± 0.0010	0.4359 ± 0.0700	0.0694 ± 0.0472	8.9102	1.3712 ± 0.0432
	2560	5	0.0584 ± 0.0511	0.4996 ± 0.0005	0.4145 ± 0.0889	0.0890 ± 0.0667	9.5116	1.4176 ± 0.0329

Table 2: Validation Set Masked Sequence In-Filling

Dataset	Hidden Size	Depth	Sequence MAE	Cosine Similarity
HPA	480	68	0.8628 ± 0.0951	0.9504 ± 0.0237
	640	55	0.7917 ± 0.1245	0.9577 ± 0.0216
	1280	25	0.6512 ± 0.1794	0.9708 ± 0.0163
	2560	5	**0.5759 ± 0.2322**	**0.9722 ± 0.0210**
OpenCell	480	68	0.7507 ± 0.1709	0.9533 ± 0.0285
	640	55	0.6641 ± 0.1764	0.9610 ± 0.0272
	1280	25	0.5698 ± 0.2016	0.9709 ± 0.0220
	2560	5	**0.4950 ± 0.2456**	**0.9711 ± 0.0271**

We also observed models had stronger performance with respect to the dataset on which they were trained. Notably, the model trained on the HPA dataset outperformed the OpenCell-trained model on the OpenCell dataset, showcasing lower PDF MAE values across all categories. This HPA model also displayed lower FID on the OpenCell validation set, underscoring the advantage of having a more extensive dataset even under differing imaging conditions. The `OpenCell_480` model achieved the best scores in half of the evaluated metrics: MAPE, MAE, SSIM, and IS.

5.2 Masked Sequence In-Filling

To test each model's sequence learning, we used a masked in-filling task similar to the training task. Similar to Section 5.1, we provide CELL-E 2 with a randomly masked (15%) sequence, an unmasked nucleus image, and an unmasked protein threshold image. To select the sequence prediction, we perform a weighted random sampling operation from the 3 amino acids with the highest predicted probabilities. We measured the accuracy as the percentage of correct predictions (noted as "Sequence MAE", see Appendix C.2). We then embedded each reconstructed sequence with `esm2_t36_3B_UR50D`, the largest model we could fit in memory, with 3B parameters, 36 layers and an embedding dimension of 2560. We computed the mean cosine similarity between the embeddings of the original and reconstructed sequences at masked positions. We show validation results in (Table 2) and all results in (Table S6).

Most models had low performance on this task in terms of reconstruction. This is understandable because the models learned to generate amino acids that were common or frequent in the dataset, but not necessarily correct for the specific sequence. On the other hand, we observed values close to 1 for the cosine similarity, indicating that the predicted amino acids had similar embedding values to the original ones at the masked positions. This could be because the models learned to capture some semantic or structural features of the amino acids, such as polarity or charge, that were reflected in the embedding space and contributed to the biological function of the sequence. Models that used the embedding model with an embedding dimension of 2560 had the best performance. For example, `OpenCEll_2560` had the best performance on both metrics, with a MAE of .4950 and cosine similarity of .9711. When compared to randomly selected amino acids for each position (Table S7), we note significantly higher Sequence MAE and Cosine Similarity.

We also note that the reconstruction ability does not improve the performance of the original language models (Table S5). This may be a result of the combined image/sequence loss used during training or a smaller corpus of data compared to datasets used for the training the original language model. Evaluation results across both datasets can be found in (Table S6).

Table 3: OpenCell Validation Set Image Prediction Accuracy after Finetuning

Fine-Tuned	Threshold Image Encoder	Nucleus Proportion MAPE	Image MAE	PDF MAE	SSIM	FID	IS
No	HPA	0.0181 ± 0.0168	0.4154 ± 0.0594	0.3887 ± 0.1270	0.1250 ± 0.1149	3.9509	2.1739 ± 0.1255
No	OpenCell	0.0161 ± 0.0148	0.4953 ± 0.0064	0.3620 ± 0.1168	0.1220 ± 0.1188	**1.5844**	2.6069 ± 0.1175
Yes	HPA	0.0166 ± 0.0151	0.3776 ± 0.0834	$\mathbf{0.3477 \pm 0.1268}$	0.1869 ± 0.1503	17.4075	2.9113 ± 0.1199
Yes	OpenCell	$\mathbf{0.0159 \pm 0.0156}$	0.4996 ± 0.0006	0.3506 ± 0.1208	0.1574 ± 0.1372	2.5026	2.7168 ± 0.1137
Yes	HPA Finetuned	0.0170 ± 0.0160	$\mathbf{0.3449 \pm 0.1305}$	0.3487 ± 0.1340	$\mathbf{0.1881 \pm 0.1541}$	19.2683	$\mathbf{3.6083 \pm 0.2013}$

5.3 Finetuning

While the HPA dataset contains information about a wide variety of proteins, the model does not innately perform as well on the OpenCell data. We considered the potential of utilizing an HPA-trained model and finetuning on the OpenCell data, thereby introducing a wider protein context than what is found in the OpenCell data alone while adapting to the imaging conditions and cell type found within the new dataset. We experimented with different finetuning strategies for CELL-E 2 on the OpenCell dataset. We used the pre-trained HPA checkpoint as the starting point for all finetuned models, continuing training on the OpenCell train set. We also evaluated the pre-trained HPA and OpenCell checkpoints without any finetuning as baselines. The finetuned models differed in how they updated the image encoders:

- `HPA Finetuned (HPA VQGAN)`: we kept the original VQGAN image encoders from the HPA checkpoint.
- `HPA Finetuned (OpenCell VQGAN)`: we replaced the image encoders with VQGANs trained only on OpenCell data.
- `HPA Finetuned (Finetuned HPA VQGAN)`: we finetuned the HPA image encoders while keeping the rest of the model frozen, then froze the image encoders and update the transformer weights.

Fig. S8 shows image predictions on an OpenCell validation protein for models with hidden size = 480. Surprisingly, the pretrained HPA model already achieved strong performance on the OpenCell dataset without any finetuning (see Table S8).

The best results were obtained by utilizing a pre-trained HPA checkpoint. We first finetuned both VQGAN image encoders while freezing the rest of the model. We then froze the VQGAN weights and allowed the base transformer to update (see Table 3). We attribute the 1.81% improvement in MAE, along with the improvements in FID and IS, to the finetuning of both the VQGANs, as it improved the consistency of image patch tokens. This provided the checkpoint with more reliable image patches to generate from. However, swapping the HPA VQGAN with an OpenCell one led to a similar losses of distribution information seen in Fig. S7. This could be because the model overfits before being able to learn probabilities across tokens. The learning obstacle comes from the possibility that images patches within the finetuned OpenCell VQGAN have sufficient (or even more) pixel consistency with the images, but the patch positional indices are misaligned with those of the HPA VQGAN. These findings are consistent with those found in analogous text-to-image works utilizing diffusion models. We did not find that finetuning improved the model's sequence reconstruction ability (see Table S9).

6 Discussion

6.1 CELL-E Comparison

In Table S10 and Table S11, we compare the performance of image localization prediction from scratch for CELL-E 2 and CELL-E. On the OpenCell validation set, CELL-E under-performs CELL-E 2 both before and after finetuning with regards to Nucleus Proportion MAPE. CELL-E 2 achieves worse Image and PDF MAE metrics before finetuning, however after finetuning CELL-E 2 achieves a 2.2% improvement for Image MAE and 1.7% for PDF MAE. On the contrary, CELL-E performs better with respect to image fidelity metrics SSIM and FID.

With respect to generation time, we found that the CELL-E 2 with hidden size of 480 was able to generate a prediction $65\times$ faster than the CELL-E model. This is a result of CELL-E 2's capability

to generate a prediction in a single step (.2784 seconds), which enables the advent of large-scale *in silico* mutagenesis studies.

6.2 *De novo* NLS Design

CELL-E 2's bidirectional integration of sequence and image information allows for an entirely novel image-based approach to *de novo* protein design. We applied CELL-E 2 to generate NLSs, which is a short amino acid sequence motif that can relocate a target protein into the cell nucleus when append to the target protein. In this case, our choice of the target protein is the Green Fluorescent Protein (GFP), a common protein engineering target [62–64] that is non-native to the human proteome and absent in the datasets. NLSs are usually identified by experimental mutagenesis studies or *in silico* screens that search for frequent sequences in nuclear proteins [51, 65]. However, these methods may yield candidates that are highly similar to known ones or not specific to the target protein. A more recent approach uses machine learning on sequence identity to augment featurization and statistical priors [17], but it is limited by the distribution of training samples due to the scarcity of experimentally verified NLSs. CELL-E 2 overcomes these limitations because it does not rely on explicit labels, and can therefore leverage significantly more unlabelled image data.

We generated a list of 255 novel NLS sequences for GFP using the procedure described in Appendix D.2. Briefly, we insert mask tokens of set length in a GFP sequence and task model with best sequence in-filling performance (`OpenCell_2560`) to fill in the masked amino acids, conditioned on a threshold image generated from the nucleus image (via Cellpose segmentation [66]). To verify the accuracy of the prediction, we pass the predicted sequence through the best performing image model (`HPA Finteuned (Finetuned HPA VQGAN)_480`), and quantify the proportion of signal intensity within the nucleus of the predicted threshold image (Fig. S9). The NLS sequences were then ranked based on sequence and embedding similarity with known NLSs (see Appendix D.2). The list of candidates can be found in Appendix D.3. We found several NLS candidates with high predicted signal in the nucleus, but which were fairly dissimilar from any protein found within NLSdb [65].

Classical NLSs are characterized by having regions of basic, positively charged amino acids arginine (R) and lysine (K) [67, 68], and are categorized as "monopartite" or "bipartite", either having a single cluster of basic amino acids or two clusters separated by a linker, respectively [69]. We observed a postive correlation between percentage of R and K residues in our predicted NLSs and sequence homology with known NLSs (Table S12). The number of clusters per sequence followed a similar trend, with sequences with relatively low sequence homology (Max ID $\% \leq 33$) having at most 2 clusters in 88% of predictions (Fig. S10). The remaining predictions, if correct, represent non-classical NLSs.

To further verify our predicted sequences, we passed the predicted NLS appended to GFP through Deeploc 2.0 [32], a leading sequence-to-class protein localization model, which predicted 89% of generated sequences were nuclear localizing and 91% contained a potential nuclear localizing signal.

Similar to CELL-E, we observed high attention weights on documented localization sequences correlated with positive protein signal within the threshold image (Fig. S11). For sequences with high predicted nucleus proportion intensities, we observed high activation across the entire sequence (novel NLS and GFP residues), with some NLS weights being an order of magnitude higher than others across the GFP sequences (Fig. 3). On the contrary, predicted sequences with comparatively less predicted intensity within the nucleus had low activation across the sequence, with little to none in the proposed NLS. We observed similar amounts of attention placed on the nucleus image patches, which largely corresponded to the location of the predicted threshold patches.

7 Conclusion & Future Work

In this paper, we have presented CELL-E 2, a novel bidirectional NAR model for protein design and engineering. CELL-E 2 can generate both image and sequence predictions, handle multimodal inputs and outputs, and run significantly faster than the state-of-the-art. By pre-training on a large HPA dataset and fine-tuning on OpenCell, CELL-E 2 can achieve competitive or superior performance on image and sequence reconstruction tasks. However, one limitation of CELL-E 2 is its output resolution, which is currently (256×256). This resolution may not capture the fine details of microscopy images, which may limit applications in real-world use where Megapixel images are ac-

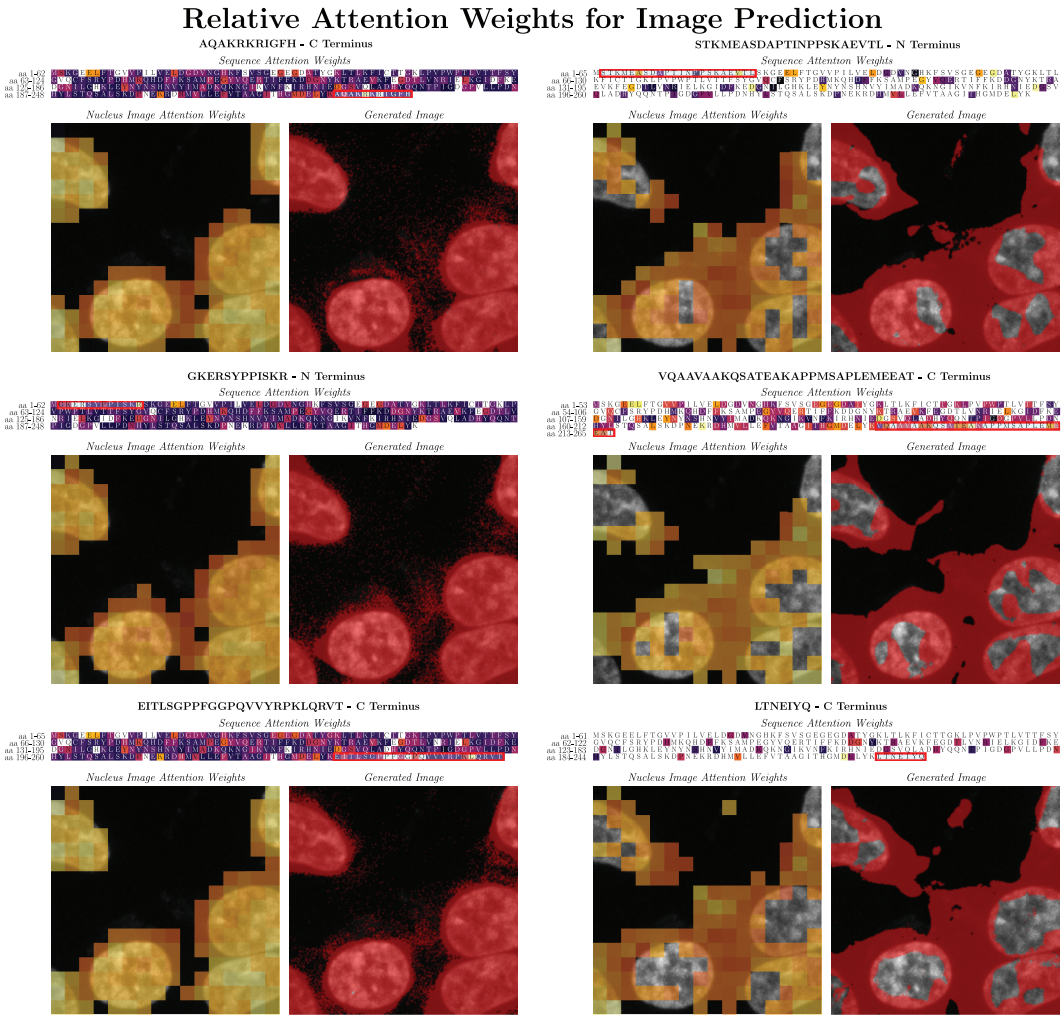

Figure 3: Attention weights associated with positive signal within the predicted image. Tokens with higher attention weight associated with background patches (low signal) are not highlighted. See Appendix D.3 for more information about the visualization process. We show 3 sequences with the highest (left column) and lowest (right column, not included in Table S13) predicted nucleus proportion intensity. The NLS+GFP sequences are shown with the designed NLS boxed in red.

quired. Increasing the output resolution of CELL-E 2 is one direction for future work. Furthermore, the sequence prediction struggles with the prediction of large stretches of amino acids as opposed to singular masked positions. Within this work, we encountered a trade-off between sequence prediction quality and prediction speed which may be overcome by reformulating the masking strategy. Similar findings were seen when compared with CELL-E, where we found accuracy measurements to improve with CELL-E 2 at the detriment of image quality metrics. The in-order prediction sequence we utilized in this paper may serve as a bottleneck for protein engineering applications despite the speed advantages gained from using a NAR architecture.

Another direction for future work is to incorporate structural information into the sequence embeddings. CELL-E 2 can generate novel NLS sequences with similar properties to GFP but low homology to existing sequences. However, the current sequence embeddings are based on a language model that may not capture all the structural features of the proteins. These features may affect the image appearance and vice versa.

We believe that CELL-E 2 is a promising model for protein design and engineering. We hope that our work will inspire more research on bidirectional NAR models for this domain and other domains that involve multimodal data.

8 Acknowledgments.

This research is supported in part by NIH grants R01GM131641 (BH) and R35-GM134922 (YSS). B.H. is a Chan Zuckerberg Biohub Investigator. A.A. contributed to visualizations and demos. E.K. was responsible for exploration and code.

References

[1] 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.

[2] Mien-Chie Hung and Wolfgang Link. Protein localization in disease and therapy. *Journal of Cell Science*, 124(Pt 20):3381–3392, October 2011. ISSN 1477-9137. doi: 10.1242/jcs.089110.

[3] 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, January 2021. ISSN 2001-0370. doi: 10.1016/j.csbj.2021.08.027. URL https://www.sciencedirect.com/science/article/pii/S2001037021003585.

[4] Wen-Yun Yang, Bao-Liang Lu, and Yang Yang. A Comparative Study on Feature Extraction from Protein Sequences for Subcellular Localization Prediction. In *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pages 1–8, Toronto, ON, Canada, September 2006. IEEE. ISBN 978-1-4244-0623-4 978-1-4244-0624-1. doi: 10.1109/CIBCB.2006.330991. URL http://ieeexplore.ieee.org/document/4133173/.

[5] Tanel Pärnamaa and Leopold Parts. Accurate Classification of Protein Subcellular Localization from High-Throughput Microscopy Images Using Deep Learning. *G3 Genes|Genomes|Genetics*, 7(5):1385–1392, May 2017. ISSN 2160-1836. doi: 10.1534/g3.116.033654. URL https://doi.org/10.1534/g3.116.033654.

[6] Sonam Aggarwal, Sheifali Gupta, and Rakesh Ahuja. A Review on Protein Subcellular Localization Prediction using Microscopic Images. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 72–77, October 2021. doi: 10.1109/ISPCC53510.2021.9609437. ISSN: 2643-8615.

[7] Yuexu Jiang, Duolin Wang, Weiwei Wang, and Dong Xu. Computational methods for protein localization prediction. *Computational and Structural Biotechnology Journal*, 19:5834–5844, January 2021. ISSN 2001-0370. doi: 10.1016/j.csbj.2021.10.023. URL https://www.sciencedirect.com/science/article/pii/S2001037021004451.

[8] Gaofeng Pan, Chao Sun, Zijun Liao, and Jijun Tang. Machine and Deep LearningDeep learning (DL) for Prediction of Subcellular Localization. In Daniela Cecconi, editor, *Proteomics Data Analysis*, pages 249–261. Springer US, New York, NY, 2021. ISBN 978-1-07-161641-3. doi: 10.1007/978-1-0716-1641-3_15. URL https://doi.org/10.1007/978-1-0716-1641-3_15.

[9] 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.

[10] Alexandra M. Schnoes, David C. Ream, Alexander W. Thorman, Patricia C. Babbitt, and Iddo Friedberg. Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space. *PLOS Computational Biology*, 9(5):e1003063, May 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003063. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003063. Publisher: Public Library of Science.

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. *arXiv:2102.12092 [cs]*, February 2021. URL http://arxiv.org/abs/2102.12092. arXiv: 2102.12092.

[12] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering Text-to-Image Generation via Transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 19822–19835. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a4d92e2cd541fca87e4620aba658316d-Abstract.html.

[13] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors, March 2022. URL http://arxiv.org/abs/2203.13131. arXiv:2203.13131 [cs].

[14] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation, June 2022. URL http://arxiv.org/abs/2206.10789. arXiv:2206.10789 [cs].

[15] Gaurav Bhardwaj, Jacob OConnor, Stephen Rettie, Yen-Hua Huang, Theresa A. Ramelot, Vikram Khipple Mulligan, Gizem Gokce Alpkilic, Jonathan Palmer, Asim K. Bera, Matthew J. Bick, Maddalena Di Piazza, Xinting Li, Parisa Hosseinzadeh, Timothy W. Craven, Roberto Tejero, Anna Lauko, Ryan Choi, Calina Glynn, Linlin Dong, Robert Griffin, Wesley C. van Voorhis, Jose Rodriguez, Lance Stewart, Gaetano T. Montelione, David Craik, and David Baker. Accurate de novo design of membrane-traversing macrocycles. *Cell*, 185(19):3520–3532.e26, September 2022. ISSN 00928674. doi: 10.1016/j.cell.2022.07.019. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867422009229.

[16] Jing Yang (John) Wang, Alena Khmelinskaia, William Sheffler, Marcos C. Miranda, Aleksandar Antanasijevic, Andrew J. Borst, Susana V. Torres, Chelsea Shu, Yang Hsia, Una Nattermann, Daniel Ellis, Carl Walkey, Maggie Ahlrichs, Sidney Chan, Alex Kang, Hannah Nguyen, Claire Sydeman, Banumathi Sankaran, Mengyu Wu, Asim K. Bera, Lauren Carter, Brooke Fiala, Michael Murphy, David Baker, Andrew B. Ward, and Neil P. King. Improving the secretion of designed protein assemblies through negative design of cryptic transmembrane domains. *Proceedings of the National Academy of Sciences*, 120(11):e2214556120, March 2023. doi: 10.1073/pnas.2214556120. URL https://www.pnas.org/doi/10.1073/pnas.2214556120. Publisher: Proceedings of the National Academy of Sciences.

[17] Yun Guo, Yang Yang, Yan Huang, and Hong-Bin Shen. Discovering nuclear targeting signal sequence through protein language learning and multivariate analysis. *Analytical Biochemistry*, 591:113565, February 2020. ISSN 0003-2697. doi: 10.1016/j.ab.2019.113565. URL https://www.sciencedirect.com/science/article/pii/S0003269719310061.

[18] 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.

[19] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers. pages 12104–12113, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhai_Scaling_Vision_Transformers_CVPR_2022_paper.html.

[20] Anthony T. Annunziato. DNA Packaging: Nucleosomes and Chromatin | Learn Science at Scitable. *DNA Packaging: Nucelosomes and Chromatin*, 1(1):26, 2008. URL http://www.nature.com/scitable/topicpage/

dna-packaging-nucleosomes-and-chromatin-310. Cg_cat: DNA Packaging: Nucleosomes and Chromatin Cg_level: MED Cg_topic: DNA Packaging: Nucleosomes and Chromatin.

[21] Peter J. Thul and Cecilia Lindskog. The human protein atlas: A spatial map of the human proteome. *Protein Science: A Publication of the Protein Society*, 27(1):233–244, January 2018. ISSN 1469-896X. doi: 10.1002/pro.3307.

[22] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE. *Advances in Neural Information Processing Systems*, 32:9689–9701, December 2019. ISSN 1049-5258.

[23] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling, January 2023. URL http://arxiv.org/abs/2301.06568. arXiv:2301.06568 [cs, q-bio].

[24] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381.

[25] 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.

[26] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model, December 2022. URL https://www.biorxiv.org/content/10.1101/2022.07.20.500902v3. Pages: 2022.07.20.500902 Section: New Results.

[27] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence, January 2022. URL http://biorxiv.org/content/early/2022/07/22/2022.07.21.500999.abstract.

[28] Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins, December 2022. URL https://www.biorxiv.org/content/10.1101/2022.12.21.521521v1. Pages: 2022.12.21.521521 Section: New Results.

[29] Henrik Nielsen, Konstantinos D. Tsirigos, Søren Brunak, and Gunnar von Heijne. A Brief History of Protein Sorting Prediction. *The Protein Journal*, 38(3):200–216, June 2019. ISSN 1875-8355. doi: 10.1007/s10930-019-09838-3. URL https://doi.org/10.1007/s10930-019-09838-3.

[30] Katelyn C Cook and Ileana M Cristea. Location is everything: protein translocations as a viral infection strategy. *Current Opinion in Chemical Biology*, 48:34–43, February 2019. ISSN 1367-5931. doi: 10.1016/j.cbpa.2018.09.021. URL https://www.sciencedirect.com/science/article/pii/S1367593118300826.

[31] Josie A. Christopher, Charlotte Stadler, Claire E. Martin, Marcel Morgenstern, Yanbo Pan, Cora N. Betsinger, David G. Rattray, Diana Mahdessian, Anne-Claude Gingras, Bettina Warscheid, Janne Lehtiö, Ileana M. Cristea, Leonard J. Foster, Andrew Emili, and Kathryn S. Lilley. Subcellular proteomics. *Nature Reviews Methods Primers*, 1(1):1–24, April 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00029-y. URL https://www.nature.com/articles/s43586-021-00029-y. Number: 1 Publisher: Nature Publishing Group.

[32] Vineet Thumuluri, José Juan AlmagroăArmenteros, AlexanderăRosenberg 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, July 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac278. URL https://doi.org/10.1093/nar/gkac278.

[33] Leyi Wei, Yijie Ding, Ran Su, Jijun Tang, and Quan Zou. Prediction of human protein subcellular localization using deep learning. *Journal of Parallel and Distributed Computing*, 117:212–217, July 2018. ISSN 0743-7315. doi: 10.1016/j.jpdc.2017.08.009. URL https://www.sciencedirect.com/science/article/pii/S0743731517302393.

[34] Bin Yu, Wenying Qiu, Cheng Chen, Anjun Ma, Jing Jiang, Hongyan Zhou, and Qin Ma. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics*, 36(4):1074–1081, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz734. URL https://doi.org/10.1093/bioinformatics/btz734.

[35] 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, November 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx431. URL https://doi.org/10.1093/bioinformatics/btx431.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022. URL http://arxiv.org/abs/2204.06125. arXiv:2204.06125 [cs].

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

[38] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, March 2022. URL http://arxiv.org/abs/2112.10741. arXiv:2112.10741 [cs].

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022. URL http://arxiv.org/abs/2112.10752. arXiv:2112.10752 [cs].

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

[41] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers, May 2022. URL http://arxiv.org/abs/2204.14217. arXiv:2204.14217 [cs].

[42] 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].

[43] 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].

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

[45] 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.

[46] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, March 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq003. URL https://doi.org/10.1093/bioinformatics/btq003.

[47] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models, November 2019. URL http://arxiv.org/abs/1908.09203. arXiv:1908.09203 [cs].

[48] IUPAC-IUB Commission on Biochemical Nomenclature A One-Letter Notation for Amino Acid Sequences13. *Journal of Biological Chemistry*, 243(13):3557–3559, July 1968. ISSN 0021-9258. doi: 10.1016/S0021-9258(19)34176-6. URL https://doi.org/10.1016/S0021-9258(19)34176-6. Publisher: Elsevier.

[49] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. pages 12873–12883, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html?ref=https://githubhelp.com.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[51] M. Cokol, R. Nair, and B. Rost. Finding nuclear localization signals. *EMBO reports*, 1(5):411–415, November 2000. ISSN 1469-221X. doi: 10.1093/embo-reports/kvd092.

[52] Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, C.J. Adams-Collier, and Kenta Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, 35(suppl_2):W585–W587, July 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm259. URL https://doi.org/10.1093/nar/gkm259.

[53] Michelle S. Scott, François-Michel Boisvert, Mark D. McDowall, Angus I. Lamond, and Geoffrey J. Barton. Characterization and prediction of protein nucleolar localization sequences. *Nucleic Acids Research*, 38(21):7388–7399, November 2010. ISSN 0305-1048. doi: 10.1093/nar/gkq653. URL https://doi.org/10.1093/nar/gkq653.

[54] Yin-Yuan Mo, Chengyi Wang, and William T. Beck. A Novel Nuclear Localization Signal in Human DNA Topoisomerase I*. *Journal of Biological Chemistry*, 275(52):41107–41113, December 2000. ISSN 0021-9258. doi: 10.1074/jbc.M003135200. URL https://www.sciencedirect.com/science/article/pii/S0021925819556435.

[55] Allison Lange, Ryan E. Mills, Christopher J. Lange, Murray Stewart, Scott E. Devine, and Anita H. Corbett. Classical Nuclear Localization Signals: Definition, Function, and Interaction with Importin *. *Journal of Biological Chemistry*, 282(8):5101–5105, February 2007. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.R600026200. URL https://www.jbc.org/article/S0021-9258(20)68801-9/abstract. Publisher: Elsevier.

[56] Ulrike Schnell, Freark Dijk, Klaas A. Sjollema, and Ben N. G. Giepmans. Immunolabeling artifacts and the need for live-cell imaging. *Nature Methods*, 9(2):152–158, February 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1855. URL https://www.nature.com/articles/nmeth.1855. Number: 2 Publisher: Nature Publishing Group.

[57] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of proteinprotein interactions. *Nature*, 417(6887):399–403, May 2002. ISSN 1476-4687. doi: 10.1038/nature750. URL https://www.nature.com/articles/nature750. Number: 6887 Publisher: Nature Publishing Group.

[58] Justin Pinkney. How to fine tune stable diffusion: how we made the text-to-pokemon model at Lambda, September 2022. URL https://lambdalabs.com/blog/how-to-fine-tune-stable-diffusion-how-we-made-the-text-to-pokemon-model-at-lambda.

[59] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, August 2022. URL http://arxiv.org/abs/2208.01618. arXiv:2208.01618 [cs].

[60] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, March 2023. URL http://arxiv.org/abs/2208.12242. arXiv:2208.12242 [cs].

[61] Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains, October 2022. URL http://arxiv.org/abs/2210.04133. arXiv:2210.04133 [cs].

[62] R. H. Köhler, W. R. Zipfel, W. W. Webb, and M. R. Hanson. The green fluorescent protein as a marker to visualize plant mitochondria in vivo. *The Plant Journal: For Cell and Molecular Biology*, 11(3):613–621, March 1997. ISSN 0960-7412. doi: 10.1046/j.1365-313x.1997.11030613.x.

[63] Nicole Maria Seibel, Jihane Eljouni, Marcus Michael Nalaskowski, and Wolfgang Hampe. Nuclear localization of enhanced green fluorescent protein homomultimers. *Analytical Biochemistry*, 368(1):95–99, September 2007. ISSN 0003-2697. doi: 10.1016/j.ab.2007.05.025. URL https://www.sciencedirect.com/science/article/pii/S0003269707003399.

[64] Akira Kitamura, Yusaku Nakayama, and Masataka Kinjo. Efficient and dynamic nuclear localization of green fluorescent protein via RNA binding. *Biochemical and Biophysical Research Communications*, 463(3):401–406, July 2015. ISSN 1090-2104. doi: 10.1016/j.bbrc.2015.05.084.

[65] 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.

[66] 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.

[67] Alex N. Nguyen Ba, Anastassia Pogoutse, Nicholas Provart, and Alan M. Moses. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics*, 10(1):202, June 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-202. URL https://doi.org/10.1186/1471-2105-10-202.

[68] 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.

[69] K. J. Bradley, M. R. Bowl, S. E. Williams, B. N. Ahmad, C. J. Partridge, A. L. Patmanidi, A. M. Kennedy, N. Y. Loh, and R. V. Thakker. Parafibromin is a nuclear protein with a functional monopartite nuclear localization signal. *Oncogene*, 26(8):1213–1221, February 2007. ISSN 0950-9232. doi: 10.1038/sj.onc.1209893.