brEMbow: Expressing Marker Proteins in EM Data

Technical Report

Esha Karlekar[1], Jan Funke[2]

1: Thomas Jefferson HSST, USA
2: HHMI Janelia, Ashburn, USA

1 Introduction

The overall goal of this project is to express marker proteins randomly in neural cells, image the cells using electron microscopy (EM), and use machine learning methods to identify cages and thus watermark cells in neural tissue.

One major problem to be solved in the field of connectomics is how to segment data effectively, especially that of EM data. The ways of piecing different EM images together are often unreliable and require copious amounts of human effort. To solve this, we must have a way of marking different neural cells so that it is easier to piece together large sets of separate EM images.

Past projects, such as Lichtman and Sanes's "Brainbow" [Livet et al., 2007], were able to mark different cells with varying combinations of colorful fluorescent proteins and image them with light microscopy. Therefore, each cell was marked with a different color (Figure 1).

We aimed to apply the same idea to electron microscopy, hence the name "brEMbow". There are many reasons to use EM data over light microscopy data, such as a higher range of magnification and a higher resolution. However, since EM data is black and white, we had to find an alternative to colorful fluorescent proteins. We used novel artificial proteins. These proteins were developed by David Baker, Deepesh Nagarajan, and Joshua Lubner from the Institute for Protein Design. His team utilized computer generated protein structures to create amino acid sequences to create synthetic proteins [Baker, 2019]. One such novel protein shape he created resembles a cage (Figure 2), and is what we used during our experiments.

The reason our cages (and other proteins developed by Baker's team) work as appropriate watermarks is because they are unlike any naturally occurring protein in cells by utilizing unique repeat proteins, sequences, and structures [Baker, 2019].

Additionally, we can be sure that these cages will not cross the cell membrane, therefore ensuring that each cell can be watermarked differently. There are certain minimum necessary components required to synthetically engineer proteins that can span membranes. A transmembrane domain fused to the N-terminus of a protein is required to move it through the endoplasmic reticulum (ER), a signal peptidase cleavage site downstream of this serves to sort it out of the ER, and an additional transmembrane domain must accumulate proteins able to span the membrane. [Qudrat and Truong, 2016] These cages were designed without such transmembrane domains and cleavage sites and will therefore remain in the cell in which they were inserted. This allows each cell to be marked distinctly.

For the purposes of electron microscopy, we chose cages that had high numbers of uranium binding sites since uranium compounds are generally used for staining. Due to the pandemic, we were not able to go into the lab to actually insert the cages into the appropriate cells, so we began our in-silico experiments by developing a way to render the cages artificially in existing fly brain CREMI EM data. Past projects have similarly used deep learning models to detect artificially rendered nanoparticles on the surface of EM data [Kharin, 2020].

We read in the Uranium binding site 3D locations from the protein cage data that David Baker's team sent us. Using this, we were able to establish how an electron microscope would actually image a cage. We rendered each cage using a Gaussian blur point spread function (as an approximation for the Airy disk model [Roels et al., 2016]) (Figure 2) and simulated random densities, orientations, and distributions in fly EM data. By filtering locations with a mask, we were able to render cage distributions only in certain segments/cells. For each cell/segment, a given probability would decide whether or not a cell had cages, and if it did, we would use a given intensity and density to render the cages with. We utilized densities varying from 1e-5 to 1e-6 and intensities varying from 0.125 to 0.025. All of this, of course, was done in such a way that the simulated cages looked like something that could appear in EM data.

Next, we developed the neural network to actually detect those cages. To do this, we developed a way to add cages on-the-fly using the Funke Lab's gunpowder library. We then built a training pipeline for a 3D U-Net. We trained several instances of U-Nets to predict the presence or absence of cages in a certain segment. We then grid-searched the hyperparameters of the method (such as density of cages, type of cage, etc) to be able to determine the minimum/maximum/ideal values of each parameter (i.e. what is the lowest density or intensity the cages can be so that the model can still effectively segment the cells). That information could then be used to educate the staining and insertion protocol when conducting the experiment using real cells and cages in the lab. So far, we have yielded very promising results, with the model being able to segment cells very accurately with low densities of

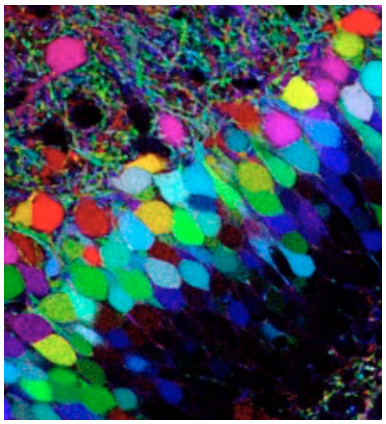

Figure 1: Cells marked using Brainbow techniques [Livet et al., 2007].

Figure 2: Example cage (Δcage01) used during the experiments reported here. Left: 3D rendering of Δcage01 using pymol. Middle: Location of uranium binding sites projected in a single EM section. Right: Rendering of the same cage under a Gaussian PSF.

cages, even in times when the human eye may not be able to detect them (Figure 3).

As stated previously, the eventual goal of this would be to actually insert the cages into neural cells and image them with an electron microscope, using the U-Net we developed previously to predict the presence or absence of cages or the presence of different types of artificial proteins. We plan to use different kinds of artificial proteins or combinations of different proteins as the parallel to different colors in “Brainbow". We currently have the ability to insert three different types of artificial proteins into zebrafish neural cells. We plan to investigate if the proteins would work as watermarks for segmenting cells and to begin testing our model on non-artificially rendered data.

2 Results

See Figure 2 for an example of a single protein cage. See Figure 3 to see an example of rendering and predictions.

References

David Baker. What has de novo protein design taught us about protein folding and biophysics? *Protein Science*, 28(4):678–683, 2019.

A Yu Kharin. Deep learning for scanning electron microscopy: Synthetic data for the nanoparticles detection. *Ultramicroscopy*, 219:113125, 2020.

Jean Livet, Tamily A. Weissman, Hyuno Kang, Ryan W. Draft, Ju Lu, Robyn A. Bennis, Joshua R. Sanes, and Jeff W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, November 2007. ISSN 1476-4687. doi: 10.1038/nature06293.

Anam Qudrat and Kevin Truong. Modular assembly of synthetic proteins that span the plasma membrane in mammalian cells. *BMC biotechnology*, 16(1):1–8, 2016.

Joris Roels, Jan Aelterman, Jonas De Vylder, Hiep Luong, Yvan Saeys, and Wilfried Philips. Bayesian deconvolution of scanning electron microscopy images using point-spread function estimation and non-local regularization. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 443–447. Ieee, 2016.

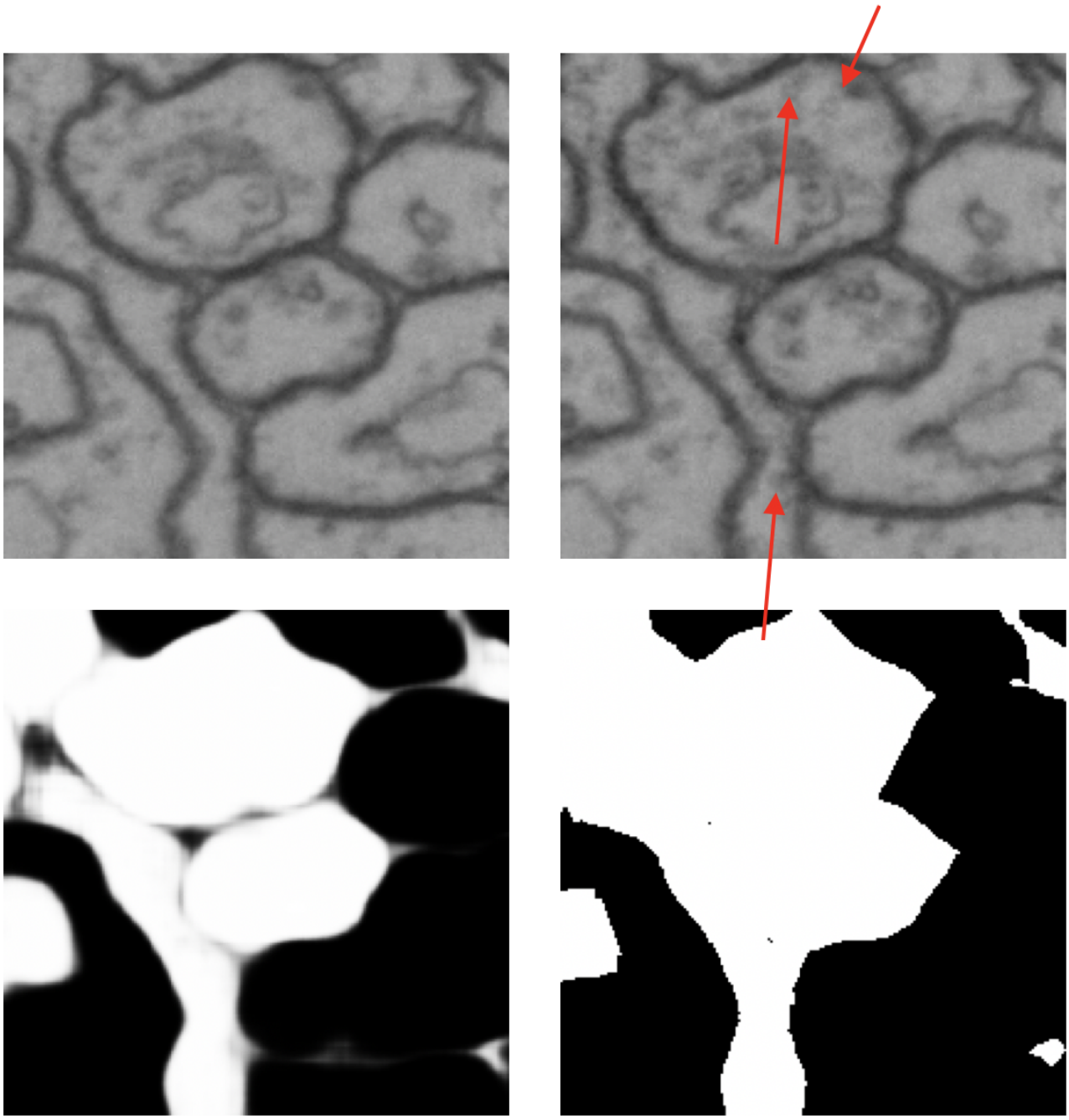

Figure 3: Top Left: EM data with no rendered cages. Top Right: EM data with rendered cages (intensity=0.075, density=1e-5), arrows pointing towards single cages. Note how it looks natural in EM data and is almost indetectable to the human eye. Bottom Left: Model prediction for this rendering, where white marks areas with cages and black marks without. Bottom Right: Ground truth for presence/absence of cages.

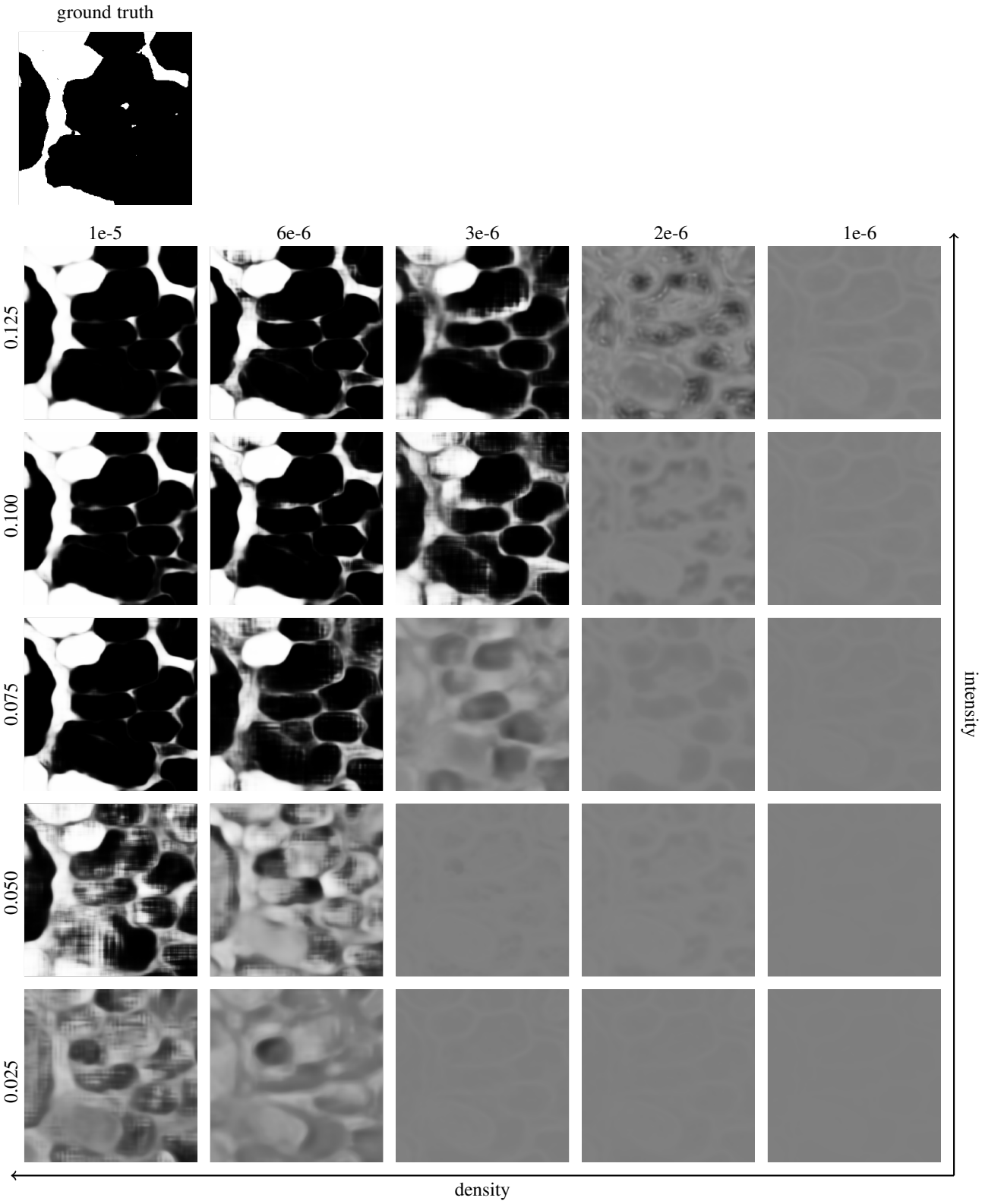

Figure 4: Prediction results for varying intensity and density values.

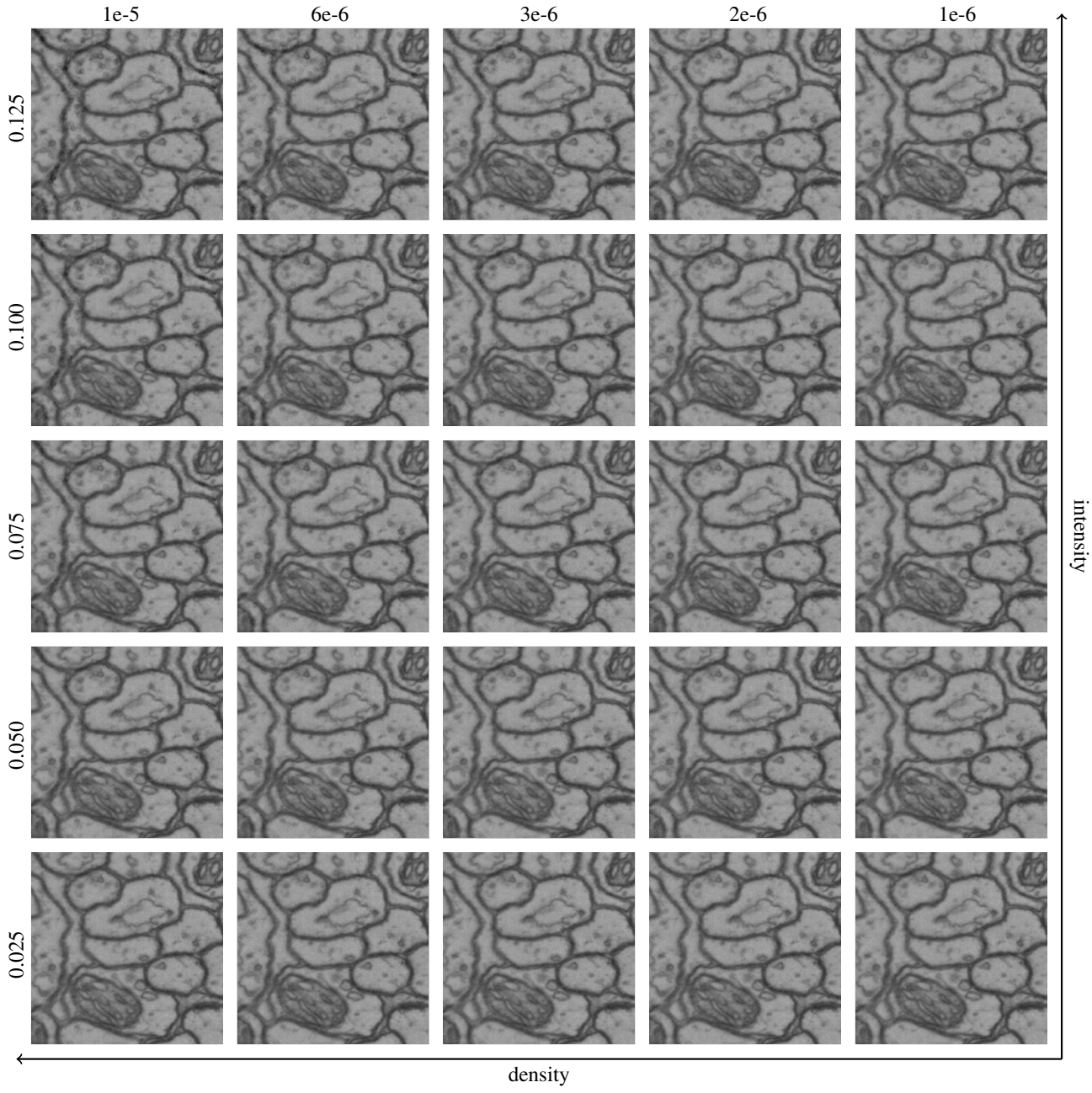

Figure 5: Rendered cages with varying intensity and density values.