# A deep generative model of 3D single-cell organization

Rory M. Donovan-Maiye[1,¤a], Jackson M. Brown[1], Caleb K. Chan[1], Liya Ding[1], Calysta Yan[1], Nathalie Gaudreault[1], Julie A. Theriot[1,2], Mary M. Maleckar[1,¤b], Theo A. Knijnenburg[1*], Gregory R. Johnson[1,¤c],

**1** Allen Institute for Cell Science, Seattle, WA, USA
**2** Department of Biology and Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

Current Addresses:
¤a Novo Nordisk, Seattle, WA, USA - the work was done prior to joining Novo Nordisk
¤b Simula Research Laboratory, Walnut Creek, CA, USA - the work was done prior to joining Simula Research Laboratory
¤c Amazon Web Services, Seattle, WA, USA - the work was done prior to joining Amazon Web Services
Contact:
* theo.knijnenburg@alleninstitute.org

## 1 Abstract

We introduce a framework for end-to-end integrative modeling of 3D single-cell multi-channel fluorescent image data of diverse subcellular structures. We employ stacked conditional $\beta$-variational autoencoders to first learn a latent representation of cell morphology, and then learn a latent representation of subcellular structure localization which is conditioned on the learned cell morphology. Our model is flexible and can be trained on images of arbitrary subcellular structures and at varying degrees of sparsity and reconstruction fidelity. We train our full model on 3D cell image data and explore design trade-offs in the 2D setting. Once trained, our model can be used to impute structures in cells where they were not imaged and to quantify the variation in the location of all subcellular structures by generating plausible instantiations of each structure in arbitrary cell geometries. We apply our trained model to a small drug perturbation screen to demonstrate its applicability to new data. We show how the latent representations of drugged cells differ from unperturbed cells as expected by on-target effects of the drugs.

## 2 Author summary

It's impossible to acquire all the information we want about every cell we're interested in in a single experiment. Even just limiting ourselves to imaging, we can only image a small set of subcellular structures in each cell. If we are interested in integrating those images into a holistic picture of cellular organization *directly from data*, there are a number of approaches one might take. Here, we leverage the fact that of the three channels we image in each cell, two stay the same across the data set; these two channels assess the cell's shape and nuclear morphology. Given these two reference channels, we learn a model of cell and nuclear morphology, and then use this as a reference frame in which to learn a representation of the localization of each subcellular

structure as measured by the third channel. We use $\beta$-variational autoencoders to learn representations of both the reference channels and representations of each subcellular structure (conditioned on the reference channels of the cell in which it was imaged). Since these models are both probabilistic and generative, we can use them to understand the variation in the data from which they were trained, to generate instantiations of new cell morphologies, and to generate imputations of structures in real cell images to create an integrated model of subcellular organization.

# 3 Introduction

Decades of biological experimentation, coupled with ever-improving advances in microscopy, have led to the identification and description of many subcellular structures in the cell that are key to cellular function. Understanding the unified role of these component structures in the context of the living cell is indeed a goal of modern-day cell biology. How do the multitude of heterogeneous subcellular structures localize in the cell, and how does this change during dynamic processes, such as the cell cycle, cell differentiation and the response to internal or environmental perturbations [1, 2]? A comprehensive understanding of global cellular organization remains challenging, and no unified model currently exists.

Advances in microscopy and live cell fluorescence imaging in particular have led to enormous insight and rich data sets with which to explore subcellular organization. However, the experimental state-of-the-art for live cell imaging is currently limited to the simultaneous visualization of only a limited number (2–6) of tagged molecules. Additionally, there are substantial, interdependent limitations regarding spatial and temporal resolution as well as duration of live cell imaging experiments. Computational approaches offer a powerful opportunity to mitigate these limitations by integrating data from diverse imaging experiments into a single model, a step toward an integrated representation of the living cell and additional insight into its function.

It should be noted that *generative models* of individual cells are particularly useful in this context, as these can capture how relationships among subcellular structures captured in experimental imaging data vary across a population of cells and encode these as distributions. Generative models may be used to understand the spatial distribution of organelles under different conditions [1, 3–5], or used as building blocks for simulations or other tasks [6, 7], depending on how these models are constructed. However, generative models also depend on both segmentation of the input microscopy images and on feature selection; results from these models are thus dependent on the quality of these segmentations and on the computed features.

Many computational models of subcellular structures are based on descriptors of texture or segmented objects extracted from microscopy images of the cell. Such image feature-based methods have previously been employed to describe and model cellular organization [8–10]. While accurate object segmentations are useful for conveying detailed information about the size, shape and localization of subcellular structures, segmentation procedures must be judiciously designed for each type of structure. Furthermore, computed image features of subcellular structures useful for e.g. generative models may not be "human-interpretable", or easily understood from a biological viewpoint. Significant amounts of effort may be spent designing features to be useful for a specific task (e.g. to measure "roughness" of a structure). Finally, ground truth data for evaluation of segmentation and feature selection can be difficult to obtain, especially for 3D cell images [11]. These challenges compound when trying to expand models to describe relationships between multiple subcellular structures and their organization within a cell.

Recent deep-learning approaches have become increasingly popular in multiple cell

biology applications and may provide an alternate computational pathway towards integrated visual representations of the cell. In cell imaging, deep neural networks (DNNs) are generally applied to perform pixel-level tasks, such as object segmentation [12], label-free prediction [13, 14], de-noising and image restoration [15, 16], or cell-level analyses such as predicting cell fates [17], classifying cell cycle status [18], distinguishing motility behaviors of different cell types [19], and subcellular pattern detection and classification [20].

The work presented here is complementary to these efforts, and provides an end-to-end framework for learning a data-driven generative model of cellular organization in a statistically principled manner. Notable in our approach is that we can build our model without the necessity of laborious hand-crafted features or subcellular segmentations.

Here we explore the use of a DNN to build a generative model deemed the Statistical Cell. This model has several applications. First, it learns *de novo* a non-parametric representation of cell and nuclear morphology. This representation is condensed into a small number of disentangled latent dimensions of morphological variation and can be used for a number of downstream tasks such as classification, interpolation, exploration of morphological variation, and generation of novel but statistically representative cellular geometries.

This representation of cell and nuclear morphology also serves as a reference frame for the model's primary task: learning representations of the localization of many independently labeled subcellular structures. This allows us to combine experiments of individual subcellular structures to predict distributions of fluorescent labels that are not directly observed together, creating a single model of integrated cell organization. This approach is distinct from other methods described above, as it can be used to learn and measure population *distributions* of cellular geometries and organelle localizations within cells, and explore their relationships to one another, as compared to prediction of an expected localization pattern in a given microscopy image.

Finally, our model is useful for practical, nonlinear dimensionality reduction for extremely high dimensional image data. This enables us to, for example, construct a statistically meaningful "average" cell from a population, determine whether a particular cell represents a common or unusual phenotype, and quantitatively measure changes in cell organization as a function of cell state (mitotic state, drug treatment, etc.).

Below, we explain how the Statistical Cell is constructed; we discuss its useful, novel contributions and provide a critical look at its current limitations.

# 4 Results

## 4.1 Statistical Cell: a variational autoencoder that models the 3D organization of subcellular structures

In this section we begin with an overview of the model, and then proceed to present its ability to model cell morphology as well as the localization of subcellular structures.

In order to jointly model the variation of all subcellular structures in our data, we engineered a stacked conditional $\beta$-variational autoencoder and trained it end-to-end on the entirety of our data. We call this model the Statistical Cell. The Statistical Cell is a data-driven probabilistic model of the organization of the human cell membrane, nuclear shape and subcellular structure localization. The diverse array of subcellular structures used here represent components that serve specific functions that may be useful for understanding cellular state.

The organizing principle of our model is that the localization of subcellular structures is meaningful only in relation to the cell geometry in which they are

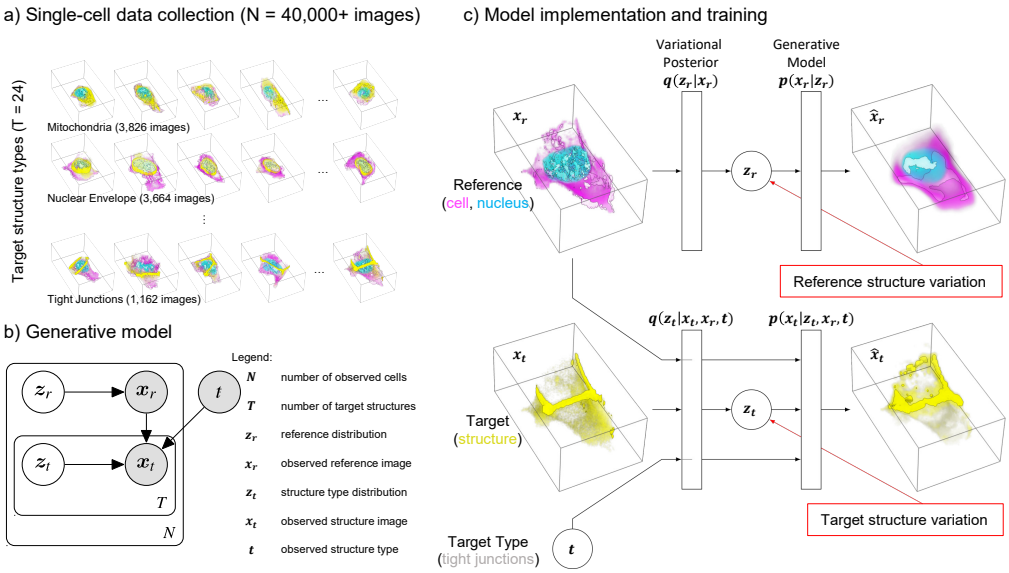



**Fig 1. The Statistical Cell.** a) A visual overview of the single-cell data collection used in this study. For each of more than 40,000 cells we have high-resolution 3D image data of the shape and location of the cell membrane (pink), nucleus (blue) and one of 24 endogenously tagged subcellular structures (yellow). The examples show actual image data of cells in the collection. b) Plate diagram of the Statistical Cell model. Shaded and un-shaded circles represent observed and learned variables, respectively. We model reference structures $\boldsymbol{x}_r$ as draws from a latent variable $\boldsymbol{z}_r$, and target structures $\boldsymbol{x}_t$ as draws from the latent variable $\boldsymbol{z}_t$, conditioned on $\boldsymbol{x}_r$ and target type $t$. c) The main components of the models are two autoencoders: one encoding the variation in the reference, i.e. the cell and nuclear shape (top), and another which learns the relationship between the target (the subcellular structures) dependent on the reference encoding (bottom).

embedded. By leveraging conditional relationships to the cell and nucleus, we allow for the integration of different subcellular structures into a single model, without these structures needing to be tagged and imaged simultaneously.

The model is trained on a collection of more than 40,000 high-resolution 3D images of live human induced pluripotent stem cells (Fig. 1a). Using 3D spinning disk confocal microscopy we collected three image channels for each of these cells: 1) Plasma membrane using CellMask Deep Red dye, 2) Nucleus via labeling DNA with NucBlue Live dye, and 3) one of 24 subcellular structures. Specifically, these cells are from clonal lines, each gene-edited to endogenously express a fluorescently tagged protein that localizes to a specific subcellular structure. To facilitate biological interpretation of our model, 5 of the 24 structures are synthetic controls, where instead of a normal GFP structure channel we present to the model either a blank channel as a structure, duplicated membrane and DNA channels, random Gaussian noise in the membrane and DNA regions, or copy a random structure channel from a different cell. There are between 1,000-4,000 cells per structure type (Table 1).

Using these images, we model cell membrane, DNA, and subcellular structure given a known target structure type (i.e. a model of cell shape, nuclear shape, and structure organization, given that the structure is one of the 24 observed structure types). This model takes the form $p(\boldsymbol{x}_{r,t}|t)$ where $\boldsymbol{x}$ is an image, $r$ indicates *reference*, i.e. the reference cell membrane and DNA dyes, $t$ indicates the *target* channel. $\boldsymbol{x}_{r,t}$ is therefore

**Table 1.** Structures used in this study, with corresponding genes, number of images and labels. Each image contains the labeled structure in addition to channels of labeled cell and nuclear shape.

| Structure Name | Gene Name (gene symbol) | # Images | Labels | |
|---|---|---|---|---|
| | | | #Interphase | #Mitosis |
| Actin filaments | actin beta (ACTB) | 2,848 | 2,544 | 304 |
| Actomyosin bundles | myosin heavy chain 10 (MYH10) | 1,392 | 1,282 | 110 |
| Adherens junctions | catenin beta 1 (CTNNB1) | 2,343 | 2,202 | 141 |
| Centrioles | centrin 2 (CETN2) | 1,605 | 1,405 | 200 |
| Desmosomes | desmoplakin (DSP) | 2,320 | 2,161 | 159 |
| Endoplasmic reticulum | SEC61 translocon beta subunit (SEC61B) | 1,120 | 1,045 | 75 |
| Endosomes | RAB5A, member RAS oncogene family (RAB5A) | 1,562 | 1,455 | 107 |
| Gap junctions | gap junction protein alpha 1 (GJA1) | 1,491 | 1,334 | 157 |
| Golgi | ST6 beta-galactoside alpha-2,6-sialyltransferase 1 (ST6GAL1) | 1,539 | 1,445 | 94 |
| Lysosomes | lysosomal associated membrane protein 1 (LAMP1) | 1,476 | 1,309 | 167 |
| Matrix adhesions | paxillin (PXN) | 1,637 | 1,531 | 106 |
| Microtubules | tubulin-alpha 1b (TUBA1B) | 2,409 | 2,219 | 190 |
| Mitochondria | translocase of outer mitochondrial membrane 20(TOMM20) | 3,826 | 3,590 | 236 |
| Nuclear envelope | lamin B1 (LMNB1) | 3,664 | 3,455 | 209 |
| Nucleoli Dense Fibrillar Component (DFC) | fibrillarin (FBL) | 1,536 | 1,407 | 129 |
| Nucleoli Granular Component (GC) | nucleophosmin 1 (NPM1) | 3,717 | 3,480 | 237 |
| Peroxisomes | solute carrier family 25 member 17 (SLC25A17) | 1,455 | 1,369 | 86 |
| Plasma membrane | Safe harbor locus (AAVS1) | 2,098 | 1,867 | 231 |
| Tight junctions | tight junction protein 1 (TJP1) | 1,162 | 1,079 | 83 |
| Control - Blank | N/A | 2,028 | 1,861 | 167 |
| Control - DNA | N/A | 2,028 | 1,861 | 167 |
| Control - Memb | N/A | 2,028 | 1,861 | 167 |
| Control - Noise | N/A | 2,028 | 1,861 | 167 |
| Control - Random | N/A | 2,028 | 1,861 | 167 |
| Total | | 49,340 | 45,484 | 3,856 |

an image of a cell containing reference structures (membrane, DNA) and a target structure (one of the 24 structure types).

Utilizing relationships in our data (Fig. 1a,b), the model is factored into two subcomponents; a reference model $\mathbf{M_R}$ that maximizes the probability of observed cell and DNA organization $p(\boldsymbol{x}_r)$, and a target model $\mathbf{M_T}$ that maximizes the conditional probability of subcellular structure organization $p(\boldsymbol{x}_t|\boldsymbol{x}_r, t)$. The complete probability model is therefore $p(\boldsymbol{x}_{r,t}|t) = p(\boldsymbol{x}_r)p(\boldsymbol{x}_t|\boldsymbol{x}_r, t)$.

Each component is modeled with a variational autoencoder (Fig. 1c) [21], allowing us to generate integrated examples from the learned data distribution, as well as map reference $\boldsymbol{x}_r$ and target $\boldsymbol{x}_t$ to learned low dimensional variables (or embeddings / latent space / latent dimensions) $\boldsymbol{z}_r$ and $\boldsymbol{z}_t$, that capture morphological variation and relationships between the reference and target structures. It is important to note that $\boldsymbol{z}_t$ is a *conditional* embedding, i.e. a low dimensional representation of information in the target image, given the information in the reference and the specific target type. For details see Section 6.2.

## 4.2 Modeling and visualizing subcellular organization via latent space embeddings

The $\beta$ variational autoencoder ($\beta$-VAE) architecture underlying the model of cell and nuclear shape ($\mathbf{M_R}$) compresses the variation in the 3D cell images into a subspace with maximal dimensionality of 512. While each cell is described by 512 coefficients in this latent space, the diagonal covariance matrix of the Gaussian prior of the latent embeddings, in tandem with the KL-divergence term in the $\beta$-VAE objective function (Eq. 3) attempts to sparsify the latent space by penalizing spurious embedding dimensions. Different values of $\beta$ lead to different numbers of effective latent dimensions; see Fig. S6 for details. (We note there is also a latent space for $\mathbf{M_T}$, but we limit our analysis and discussion in this section to the latent space of the reference model $\mathbf{M_R}$.)

We investigated the interpretability of the latent space by correlating the cell

coefficients for each latent dimension with measured cell metrics, such as cell size and mitotic state. We observed various strong correlations between cell metrics and those latent dimensions that showed a significant variation; see Fig. S3 for details. As shown in Fig. 2, the two top dimensions of the latent space visually stratify the cells based on mitotic stages as well as cell height. Importantly, cell images that are generated along these latent space dimensions show the expected phenotype (Fig. 2c,d). It is worthwhile to note that the model captures both biologically interpretable features, such as cell height, as well as features that are properties of the image measurement such as the overall fluorescent dye intensity.

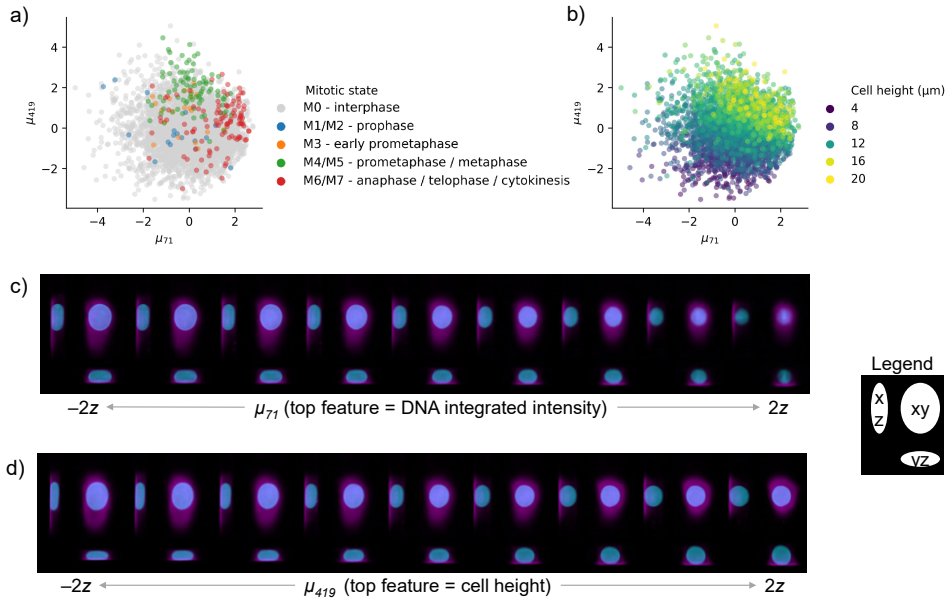



**Fig 2.** The reference latent space learned by the model encodes interpretable features and stratifies cells by biologically relevant features. a) Cells in the test set undergoing mitosis are stratified by the top two dimensions of the reference model latent space. b) The top two dimensions of the reference latent space, as ranked by mean absolute deviation, for all cells in the test set, colored by cell height. c) Maximum intensity projections of 9 generated 3d cell images along latent space dimension $\mu_{71}$. The integrated intensity of the DNA channel correlates strongly with $\mu_{71}$. d) Similar to c) but not showing latent space dimension $\mu_{419}$ which correlated strongly with cell height. The latent walks in parts c) and d) occur in nine steps that span -2 to 2 deviations of that latent dimension's variation. See Supp. Figs. S1 and S2 for more visual associations between latent dimensions and features, and Supp. Fig. S3 for exhaustive correlations of features with latent dimensions.

## 4.3 Sparsity/Reconstruction Trade-Off

Since our model is composed of $\beta$-VAEs, the $\beta$ parameter presents an important trade-off between compact representation and high fidelity image reconstruction. Tuning $\beta$ allows the modeler to preferentially weight the two components of the loss function. A high value of $\beta$ favors a compact representation of the cell in the low dimensional embedding space (i.e. the number of dimensions needed to describe an image $\boldsymbol{z}_r$ and $\boldsymbol{z}_t$) via a higher emphasis on the $KL(q(z|x)|p(z)))$ loss term. A low value of $\beta$ emphasizes accurate reconstruction of the original image by placing more weight on

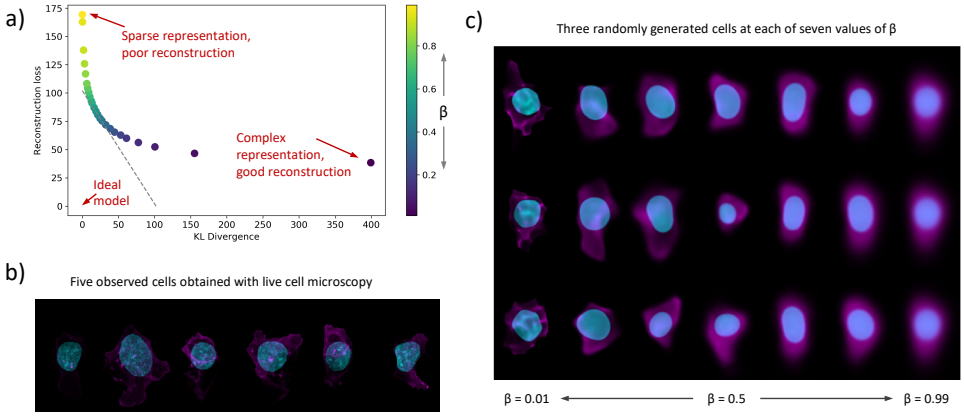



**Fig 3. Evaluation of the sparsity-reconstruction trade-off with 2D Statistical Cell models.** a) shows the average rate and distortion terms of the ELBO for models at a different trade-off $\beta$. Gray dotted line indicates the trade-off achieved by the best performing model. b) Images of six observed cells. Nucleus is in blue. Cell membrane is in magenta. c) Images of cells generated with the Statistical Cell for different values of $\beta$. See Supp. Fig. S6 for more data on model sparsity.

the $\mathbb{E}_{q(z|x)}[\log p(x|z)]$ loss term. The relative emphasis of one term versus the other has consequences for the model and for its applications. For example, one might desire a less-complex data embedding to facilitate the statistical exploration and interpretation of the latent space, while in other circumstances it might be preferable to obtain a more complex embedding that enables the comparative analysis of high-fidelity generated images.

Several methods have been proposed to modulate the trade-off between sparsity and reconstruction in VAEs [22, 23], and other factors such as data normalization, model architecture, hyper-parameters (including training schedules [24]) may also impact this relationship. To demonstrate how our model performs as a function of this relationship, we adopt a reparameterized variational objective that allows us to tune the relative weights of these two terms (see Eq. 3).

Because parameter exploration using the full 3D model is prohibitively expensive, we explored the effect of the $\beta$ parameter using a 2D model. This model is the same in all regards to the 3D model, other than that the inputs and convolutions are two dimensional instead of three. Our 2D input data was generated from the 3D data by taking a maximum-intensity projection along the $z$-axis. While this reduction obfuscates some details of the cell's structure and organization, it retains a largely faithful picture of overall cell shape and reconstruction detail, and allows us to explore model and parameter choices approximately an order of magnitude more quickly than using the 3D model.

Using the 2D data, we trained 25 models of cell and nuclear shape ($\mathbf{M_R}$) with $\beta$ values evenly spaced between 0 and 1 using 2D maximum-intensity-projected images of cell and nuclear shape. Using the test data, i.e. data not used in the training phase, we recorded the average of the two terms of ELBO for each of the 25 models and plot the two as a function of $\beta$ in Fig. 3a. Sampling from the cell and nuclear shape representation, we show generated images across a range of $\beta$ values in Fig. 3c. A few observed cells are visualized in Fig. 3b for reference. We note that compared to observed cell and nuclear shapes, generated images close to $\beta \to 0$ retain more detail and perhaps more diversity than images at $\beta \to 1$, although this comes at a trade-off of

increased representation dimensionality.

#### 4.3.1 Visualization of generated cells and conditionally generated structures

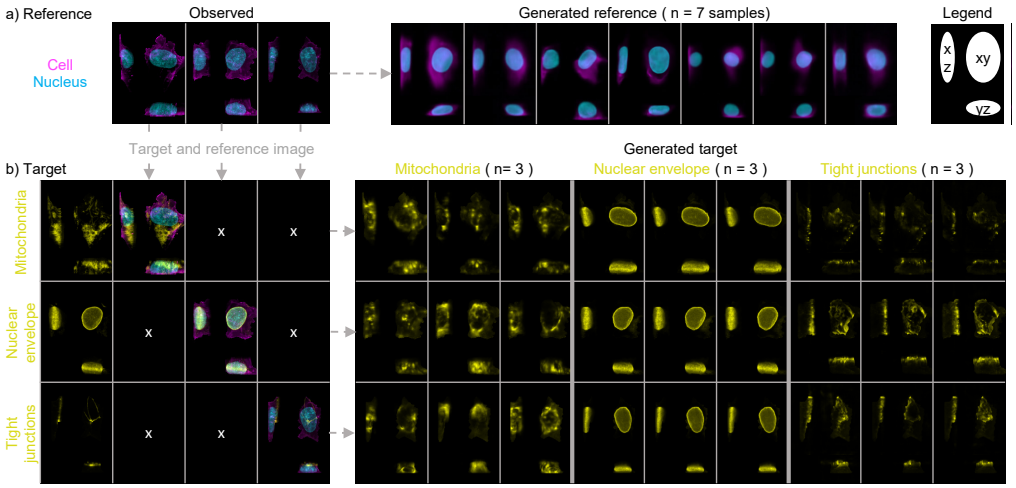



**Fig 4. Generating cell images from the Statistical Cell model**. a) Realistic images can be generated from the probabilistic model of the cell membrane and the nucleus. On the left are three examples of actual cells. On the right are seven generated cells, sampled using independent random draws from the latent space. b) Cell and nuclear image data of actual cells can be used to generate a simulated image of a given structure even if that structure was not measured for that cell. On the left are three actual cells for which three different structures were imaged. From top to bottom: mitochondria, nuclear envelope and tight junctions. On the right are depictions of the generated structure channels for each of these cells and structures. Three examples of each structure in each cell are shown, each generated using independent random draws from the latent space. See Supp. Fig. S4 and Supp. Fig. S5 for more examples.

An important application of the trained Statistical Cell model is to visualize cellular structures by generating images sampled from the probabilistic models. That is, by sampling from the latent space that describes the trained probabilistic model of structures dependent on cell and nuclear shape ($\mathbf{M_T}$), we can visualize the location and shape of subcellular structures, and how those properties vary in the data. Moreover, the construction of the model enables us to impute and visualize multiple subcellular structures in the same cell geometry simultaneously, whereas the data the model is trained on only contains one structure tagged per image. Because the model is probabilistic, it approximates the diversity of structure localization specific to each structure type. Fig. 4b depicts multiple examples of different subcellular localization patterns given an observed cell and nuclear shape.

Overall, we find that in distribution the generated structures vary in localization as one would expect, providing strong evidence that the network is successfully learning appropriate rules governing structure-specific localization that are not explicitly encoded in the images themselves or in the "target type" input. For example,

- Mitochondria are distributed throughout the cytoplasm, but are never found inside the nucleus.
- The nuclear envelope forms a closed shell around the DNA.

- The tight junctions are at the apical surface of the cell and around the cell periphery.
- The nucleoli (both the Dense Fibrillar Center and the Granular Component) form blobs that are always inside of the nucleus, never outside.
- Matrix adhesions are always at the basal surface.

The latter two examples along with other examples of typical organelle localization are found in Supp. Fig. S5. It is important to note that due to the limitations of the data and the specific model construction, generated subcellular structures are independent of each other (e.g. generated tight junctions may overlap with generated mitochondria).

## 4.4 Quantification of the coupling of subcellular structure localization to gross cellular morphology

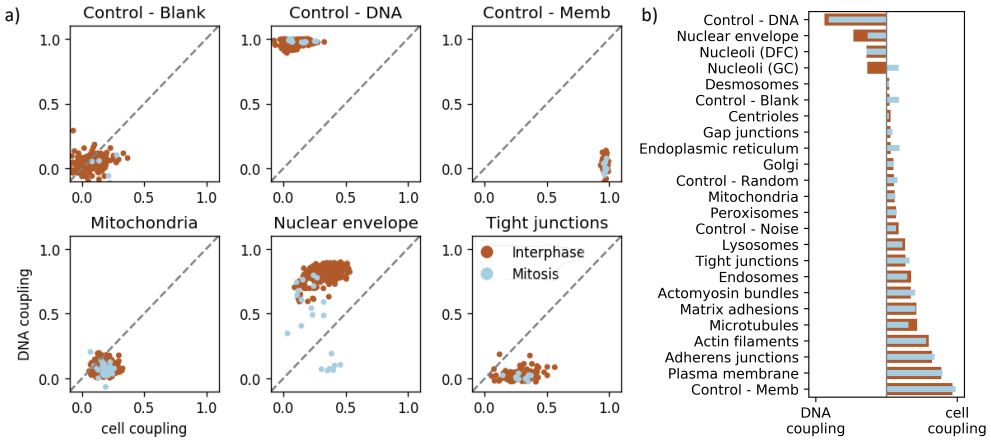



**Fig 5.** Quantification of the coupling of cellular morphology and subcellular structure. a) shows the relative coupling strength of three structures to the nuclear shape (y-axis) and cell membrane (x-axis) of the cells in which they reside, according to Eq. 4. Each point represents a cell; brown points are cells in interphase, blue points are cells undergoing mitosis. b) shows the relative degree of coupling of each structure to the cell membrane or nuclear channel, and how this changes during mitosis.

In the previous sections we aimed to show, both qualitatively and quantitatively, that the Statistical Cell enables us to model the organization of subcellular structures by leveraging the reference channels, i.e. the cell membrane and the nuclear shape. An important next question is to what extent the reference channels by themselves inform the prediction of subcellular structure organization.

To answer this question, we constructed a measure of coupling sensitivity between a subcellular structure and the morphology of either the cell membrane or the nucleus. Specifically, we quantified the sensitivity of our model to the coupling between each subcellular structure and a reference channel (say, cell membrane) by comparing the ELBO for that unperturbed image with the ELBO of a perturbed version of that image, where the reference channel (e.g. membrane) is replaced by a randomly selected membrane channel from the other cells in the population (see Methods, Eq. 4).

As controls for this metric, we created artificial "structure" channels that are duplicates of each of our reference channels (cell membrane and nucleus) to confirm that this measure of coupling makes sense in the limit of structures that are perfectly

correlated with one of the reference channels. Fig. 5a shows the coupling metrics across three subcellular structures as well the controls. We observed that each control is quantified as being coupled only to itself, and not to the other reference channel; that is, it appears that the cell membrane and the nucleus are not informative of each other under our model. We also used a blank structure as a control, and as expected did not see any coupling between it and either reference channel. Regarding the three subcellular structures: We observed that the mitochondria have low coupling to both the nucleus and cell membrane, indicating that under our model, the placement of the mitochondria is not coupled to a specific cell's morphology. On the other hand, the tight junctions, while effectively uncoupled to the nuclear channel, show a strong coupling with the cell membrane. This is to be expected, since tight junctions co-localize with the membrane at the apical perimeter of the cell. Lastly, we see that the nuclear envelope is predominately coupled to the DNA channel, except during mitosis when it dissolves, which is also as expected. More details on this are given below.

In order to understand how strongly each structure is coupled to the cell or nuclear reference structure overall, in Fig. 5b we condensed the coupling metrics across many cells, as displayed in Fig. 5a, to a single summary statistic for each of the 24 subcellular structures. The statistic that we employed is the normalized difference of the nuclear coupling to the cell membrane coupling from Eq. 4, averaged over all cells in that structure/phase of the cell cycle (see Eq. 5 for details).

Fig. 5b shows a spectrum of differential couplings under our model, all of which are biologically plausible. They range from the nuclear envelope and other nuclear tags being coupled to the nucleus, to microtubules and other cytoplasmic structures being strongly coupled to the cell membrane.

Interestingly, we observed there is differential coupling between mitotic cells (plotted in blue) and interphase cells (plotted in brown) for some subcellular structures. That is, a certain subcellular structure may be strongly coupled to the nuclear reference channel during interphase, but that coupling may breakdown during mitosis, or vice versa. For example, the nuclear channel of the mitotic cells shows a weaker coupling to the nuclear envelope than the interphase cells (Fig. 5a). Our interpretation is that the break down of the nuclear envelope during mitosis inhibits the model's ability to leverage the nuclear channel for prediction.

Fig. 5b highlights several other subcellular structures whose differential coupling (to the cell membrane vs. the nucleus) is sensitive to the position of the cell in the cell cycle. Of particular note is that we detect that the nucleolus granular component is tightly coupled to the nucleus during interphase, but that relationship is inverted during mitosis.

## 4.5 Out of sample data using drug perturbations of subcellular structures

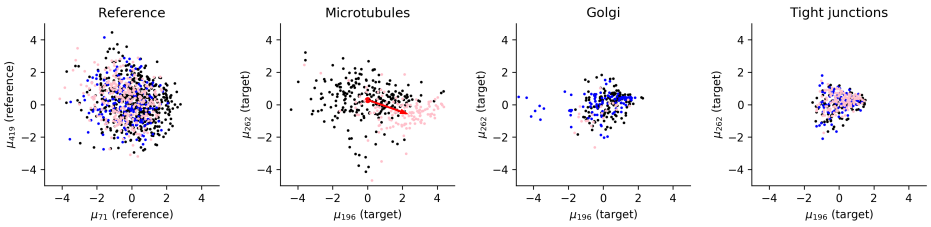
To determine the model's utility on out-of-sample data, we explore the model's ability to detect perturbations of cell morphology and subcellular localization patterns using a small drug perturbation experiment. The drug data set consists of images of cells that were fluorescently tagged with one of three subcellular structures: microtubules, Golgi, or tight junctions, and exposed to one of two drugs: Brefeldin (5.0 µM) or Paclitaxel (5.0 µM). See Fig. 6a and Section 6.1.4. This data set also contained cell images for untreated cells. The primary focus of this follow-up study was to determine whether the trained model can sensibly be applied to these drug data and provide biological insights by evaluating on low-dimensional representations of reference and target structures.

Latent embeddings (reference $\boldsymbol{z}_r$ and target $\boldsymbol{z}_t$) were computed for each image of untreated and drug-treated cells. Fig. 6b shows the first two dimensions (as ranked by

a)

| Cell + Nucleus | Microtubules | Golgi | Tight junctions |
|---|---|---|---|
| Untreated | N = 242 | N = 144 | N = 105 |
| Brefeldin 5.0 µM | N/A | N = 105 | N = 94 |
| Paclitaxel 5.0 µM | N = 132 | N = 23 | N = 138 |

b)




c)

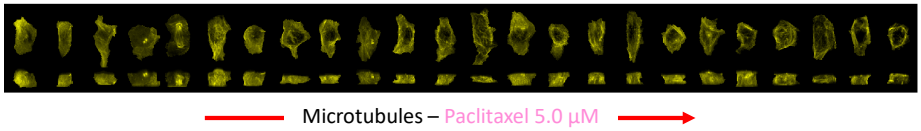



**Fig 6.** a) Experimental overview of drug-treated cells (rows) and measured subcellular structures (columns) where N indicates the number of obtained cells. Drugs are color coded (Brefeldin is blue, Paclitaxel is pink). b) Scatter plots of cells embedded in first two dimensions of the latent space of the Statistical Cell model. From left to right: Reference structures (cell + nucleus) in the reference latent space, microtubules, Golgi, and tight junctions in the conditional latent space. The red arrow in the microtubules subplot indicates the direction of the change of the centroid of the untreated (black) and treated (pink) populations. c) Visualization of the microtubules of 24 cells. For each cell we have visualized the maximum intensity projection along z, i.e. looking at the xy plane (top), and the maximum intensity projection along x, i.e. looking at the yz plane (bottom). The cells from both populations are sampled at fixed intervals from the untreated(left)-to-treated(right) direction after projecting onto the red line in b2.

mean absolute deviation) of this embedding for the reference channels (cell membrane and nucleus) and the three structures. We observe that the treated and untreated cells occupy the same space for the reference model $\mathbf{M_R}$, that is, the embeddings of the cell membrane and nuclear morphology are largely unperturbed by drug treatment in these two leading latent dimensions. Specifically, the treated (pink and blue) and untreated (black) cells do not form different clusters but are highly overlapping. Note that the structure model is conditional on the reference (cell membrane and nucleus), and if the treated cells were substantially different from the untreated cells with regards to their cell and nuclear morphology, we could not sensibly apply the conditional model.

We focus on the most visually striking and largest perturbative effect that the model detected, and here visualize the effect of Paclitaxel on microtubules by plotting the direction-of-effect on the first two dimensions of the latent representation Fig. 6b (microtubules). By sampling images uniformly across this vector (Fig. 6c), moving from left to right we see that this direction corresponds with change from uniform to condensed localization.

# 5 Discussion

The Statistical Cell is a model of the organization of subcellular structures in human induced pluripotent stem cells. We have described the model and its capabilities, and explored how it may be a useful tool for studying the organization of cells from fluorescent 3D spinning disk confocal images. In this section, we discuss important considerations and future work of the Statistical Cell.

Previous incarnations of our Statistical Cell model [25, 26] featured an adversarial loss function. Along with non-generative direct image transformation methods, e.g. [13, 15], generative adversarial networks (GANs) have been proposed to produce high-fidelity generated cell images [27, 28]. Although these adversarial-loss based methods may produce crisp images when appropriately tuned, they remain notorious in their difficulty to efficiently optimize [29, 30], suffer from several poorly understood and difficult to diagnose pathologies such as "mode collapse" [31] and difficulty to initialize models that produce large images [29, 30].

It should be noted that both the VAE loss function used here and adversarial autoencoders [32] allow for the construction of a low dimensional representations of specified distribution (e.g. Gaussian-distributed latent spaces). We ultimately chose a VAE-based model version for a number of reasons, both practical and theoretical. Compared to GANs, $\beta$-VAEs are more stable and easier to train, possess stronger theoretical bounds [21], and are able to easily trade off latent-space dimensionality for reconstruction fidelity [22] while also disentangling latent dimensions.

In this work we employ a standard conditional $\beta$-VAE architecture, composed of convolutional/residual blocks. We present no drastic architectural innovations, but rather leverage the proven reliability of $\beta$-VAEs in other domains to build a solid framework for integrating and modeling 3D spinning disk confocal fluorescent microscopy data. Our focus is the cell biological application, and the way that the $\beta$-VAE enables hands-off quantification of image data that are difficult to describe in anything other than qualitative terms.

There are a number of ways in which our model could be extended in the future. Recent work in non-biological domains has suggested that using image features and auxiliary loss functions for VAEs can improve model performance and generalizability [33]. Directly optimizing our model to retain specific morphological properties, e.g. using feature-based losses such as cell volume, cell height, etc., may improve image generation while retaining specific statistical properties relevant to the application of these models.

As in many bioimage informatics applications, image pre-processing and normalization are important issues and can have important consequences to downstream workflows. In the Statistical Cell model, image normalization and the loss function (ELBO) are coupled to each other. Changing the way in which the cell images are pre-processed directly impacts how the loss function affects the training of the model. For instance, noise and bright spots in the cell and nuclear dyes are penalized disproportionately to their biological significance. We have not thoroughly studied different normalization schemes and their effects on the inferred model, and as with many image analysis methods, it is likely that results may vary as a function of image pre-processing.

One avenue of approach to normalization could be to use image segmentations as inputs to the model, rather than the fluorescent dye intensities. Segmentation itself presents its own problems, but as 3D segmentation techniques for densely packed spinning disk confocal fluorescent microscopy images improve [11], this approach should become more feasible and deserving of serious consideration. Our framework is amenable to incorporating both segmentation-based inputs and hand-crafted features, a

future direction that holds promise for incorporating the most desirable facets of each of these approaches.

An important aspect of the Statistical Cell model is the application of the trained model to external data. For any model, it is important to understand to what extent it may be applied in new contexts. The results presented here were trained on a data set of relatively few conditions, and we do not expect our model to generate realistic images or produce accurate representations of cellular structures on images that were collected in a substantially different way, e.g. different cell types, different microscopes, etc., although future models may be able to account for these types of biological and technical variation [34].

It will be important to understand whether the Statistical Cell can, at least, be used as a pre-trained model to enable faster training on smaller data sets of cell images that were obtained under different conditions. The computational complexity of training (from scratch) a 3D Statistical Cell model can be prohibitive without access to expensive GPU resources. Specifically, training the 3D Statistical Cell model on a sizable corpus of relatively large 3D images as explained in this work required 2 weeks of training time using two high-memory (32GB) GPUs.

The Statistical Cell can visualize multiple subcellular structures simultaneously at a single-cell level. Yet, some complex interrelationships between structures may not be accurately represented by the model. This is because the Statistical Cell model captures the partial correlation between a target subcellular structure and the cell and nucleus, but not directly between different subcellular structures. Cell lines that have been gene-edited to express multiple tagged subcellular structures could be leveraged to model a richer covariation structure. Beyond jointly modeling more simultaneously acquired image channels, additional data types (e.g. RNA FISH) could be incorporated into a generative single-cell model. Conditional modeling and visualization of multiple jointly acquired data modalities is necessary to move towards a truly integrated picture of cell state.

# 6 Materials and methods

## 6.1 Data

Our model was trained on publicly available cell image data generated at the Allen Institute for Cell Science.

### 6.1.1 Allen Cell Collection

The input data for training can be obtained at `allencell.org` or directly at `https://open.quiltdata.com/b/allencell/packages/aics/pipeline_integrated_single_cell`. Imaging and culture conditions are described in [13]. Each source image consists of channels corresponding to the reference nuclear signal and cell membrane signal, and a fluorescently labeled target sub-cellular structure. Extensive information can be found on `allencell.org`.

### 6.1.2 Preprocessing of images

All cell regions were segmented from the field of view using a segmentation workflow. See `https://www.allencell.org/extracting-information.html` for more details. Each channel was processed by subtracting the most populous pixel intensity, zeroing-out negative-valued pixels, and re-scaling image intensity to a value between 0 and 1. The cells were aligned by the major axis of the cell shape, centered according to the center of mass of the segmented nuclear region, and flipped according to image skew.

**Table 2.** Drug treatments.

| Perturbation agent | Vendor | Catalog | Well-known action | Concentration | Incubation |
|---|---|---|---|---|---|
| Brefeldin A | Selleckchem.com | No.S7046 | Vesicle trafficking inhibitor | 5 µM | 0.5 hr |
| Paclitaxel | Selleckchem.com | No.S1150 | Microtubule polymer stabilizer | 5 µM | 2 hr |
| DMSO | Sigma Aldrich | N/A | Vehicle control | 0.01% | 0.5 to 2 hr |

Each of the 49,340 cell images were linearly rescaled to cubic voxels of 0.317 µm/px, and padded to $128 \times 96 \times 64$ cubic voxels. 2D images for Section 4.3 were created by maximum-intensity projection along the $z$-axis, independently for all three image channels, and are available in the data package.

### 6.1.3 Mitotic annotations

Cells were annotated as being in interphase or mitosis via manual inspection of images by a resident expert cell biologist and released as part of the Allen Institute - Integrated Mitotic Stem Cell. Mitotic cells were further annotated by which phase of mitosis they were in: prophase (M1+M2), early prometaphase (M3), prometaphase / metaphase (M4+M5), and anaphase / telophase / cytokinesis (M6+M7). Further details are available at `https://www.allencell.org/hips-cells-during-mitosis.html#sectionMethods-for-mitotic-cells`

### 6.1.4 Drug perturbations

The data for the drug perturbation experiments were acquired independently of the main training and testing data for the model, and here we describe its acquisition and processing.

A subset of mEGFP-tagged human induced pluripotent stem cells (hiPSCs) from the Allen Cell Collection were selected in this study, including TUBA1B line:AICS-12, ST6GAL1 line:AICS-25, TJP1 line:AICS-23, to show the location of a particular cell organelle or structure and represent cellular organization. mEGFP-tagged hiPSCs were seeded onto Matrigel-coated 96-well plates at a density of 2,500 to 3,500 cells per well and maintained in culture in phenol-red free mTeSR1 media with 1% penicillin streptomycin for 4 days before imaging (media changed every day). On day 4, cells on the 96-well plate were treated with one of the pre-selected, well-characterized drugs with concentration and incubation time described in Table 2.

At the end of the incubation time, hiPSCs were then incubated in imaging media of phenol red-free mTeSR1 media (Stem Cell Technologies) with 1% penicillin streptomycin (Thermo Fisher) with X1 Nuc Blue Live (Hoechst 33342, Thermo Fisher) for 20 min and 3X CellMask (Thermo Fisher) for 10 min. The cells were washed with fresh imaging media prior to being imaged live at high magnification in 3D.

3D Live-cell imaging of mEGFP-tagged hiPSCs was performed on a Zeiss spinning-disk microscope with a 100x/1.2 NA W C-Apochromat Korr UV-vis IR objective (Zeiss) and a 1.2x tube lens adapter for a final magnification of 120x, a CSU-x1 spinning-disk head (Yokogawa) and Orca Flash 4.0 camera (Hamamatsu) (pixel size 0.271 $\mu$m in x-y after 2x2 binning and 0.29 $\mu$m in Z). Standard laser lines (405, 488, 561, 640 nm), primary dichroic (RQFT 405, 488, 568, 647 nm) and the following Band Pass (BP) filter sets (Chroma) were used for fluorescent imaging: 450/50 nm for detection of Nuc Blue Live, 525/50 nm for detection of mEGFP, and 690/50 nm for detection of CMDR dye (Thermo # C10046). Cells were imaged in phenol red-free mTeSR1 media, within an incubation chamber maintaining 37°C and 5% $CO_2$. Bright field images were acquired using a white light LED with broadband peak emission using

the mEGFP BP filter 525/50 nm for bright field light collection.

To obtain single-cell measurements, fluorescent images of cells treated with CellMask Deep Red dye and NucBlue Live dye were segmented with the aforementioned segmentation workflow. The fluorescent images are normalized with a median filter and adaptive local normalization. Nuclei are segmented with active contouring. Segmented nuclei are used to create seeds for segmentation of individual cells based on the signal from the plasma membrane. The plasma membrane signal is boosted at the top of the cell and fluorescent endocytic vesicles are removed from normalized cell membrane image and are then segmented with 3D watershed with seeds from nucleus segmentation.

Further details are available at `https://www.allencell.org/drug-perturbation-pilot.html`. The data are available at `https://www.allencell.org/data-downloading.html#sectionDrugSignatureData`

## 6.2 Model Architecture

At its core, the Statistical Cell is a probabilistic model of cell and nuclear shape conjoined to a probability distribution for the localization of a given subcellular structure conditional on cell and nuclear shape. An observed image, $\boldsymbol{x}_{r,t}$, is modeled as $p(\boldsymbol{x}_{r,t}|t) = p(\boldsymbol{x}_r)p(\boldsymbol{x}_t|\boldsymbol{x}_r, t)$, where $r$ indicates *reference* image channels that contain the same cellular structure across all images (in this case the plasma membrane using CellMask Deep Red dye, and the nucleus via labeling DNA with NucBlue Live dye), and $t$ indicates *structure* channels. The model maps $\boldsymbol{x}_r$ and $\boldsymbol{x}_t$ to learn low dimensional variables (or "embeddings"), $\boldsymbol{z}_r$ and $\boldsymbol{z}_t$ that capture morphological variation and relationships between the reference and target structures, allows for sampling of missing data and realistic exemplars, and characterizes statistical relationships between the reference and target structures.

The Statistical Cell model consists of two sub-models that are trained (conditionally) independently, the model of cell and nuclear shape, $\mathbf{M_R}$, and the conditional model of structure localization, $\mathbf{M_T}$.

**Table 3.** Architecture of model used in this study. Arrows indicate spatial downsampling or upsampling.

| Reference Model, $\mathbf{M_R}$ | Target Model, $\mathbf{M_T}$ |
|---|---|
| $\boldsymbol{x}_r$ | $\boldsymbol{x}_t$ |
| Residual layer, $\downarrow$, 32 | Residual layer + $(\boldsymbol{x}_r, t)$, $\downarrow$, 32 |
| Residual layer, $\downarrow$, 64 | Residual layer + $(\boldsymbol{x}_r, t)$, $\downarrow$, 64 |
| Residual layer, $\downarrow$, 128 | Residual layer + $(\boldsymbol{x}_r, t)$, $\downarrow$, 128 |
| Residual layer, $\downarrow$, 256 | Residual layer + $(\boldsymbol{x}_r, t)$, $\downarrow$, 256 |
| Residual layer, $\downarrow$, 512 | Residual layer + $(\boldsymbol{x}_r, t)$, $\downarrow$, 512 |
| [FC 512], [FC 512] | [FC 512], [FC 512] |
| $\boldsymbol{z}_r = N(\mu_{\boldsymbol{z}_r}, \sigma_{\boldsymbol{z}_r})$ | $\boldsymbol{z}_t = N(\mu_{\boldsymbol{z}_t}, \sigma_{\boldsymbol{z}_t})$ |
| FC 512 | FC 512 |
| Residual layer, $\uparrow$, 512 | Residual layer + $(\boldsymbol{x}_r, t)$, $\uparrow$, 512 |
| Residual layer, $\uparrow$, 256 | Residual layer + $(\boldsymbol{x}_r, t)$, $\uparrow$, 256 |
| Residual layer, $\uparrow$, 128 | Residual layer + $(\boldsymbol{x}_r, t)$, $\uparrow$, 128 |
| Residual layer, $\uparrow$, 64 | Residual layer + $(\boldsymbol{x}_r, t)$, $\uparrow$, 64 |
| Residual layer, $\uparrow$, 32 | Residual layer + $(\boldsymbol{x}_r, t)$, $\uparrow$, 32 |
| Residual layer, $\uparrow$, 2 | Residual layer + $(\boldsymbol{x}_r, t)$, $\uparrow$, 1 |
| $\hat{\boldsymbol{x}}_r$ | $\hat{\boldsymbol{x}}_t$ |

A diagram of the model is shown in Fig. 1b,c. The reference component $\mathbf{M_R}$

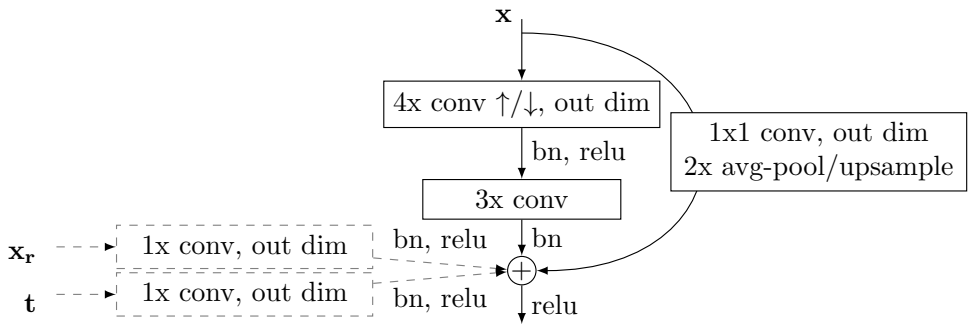



**Fig 7. Residual block used in this model.** Each layer of our model is a modified residual layer. In the encoder, the layer input, $\mathbf{x}$, is passed through a 4x convolution kernel with a stride of 2, then a 3x convolution kernel with a stride of 1 or a 1x convolution kernel with a subsequent avg-pooling step. These results are summed along the channel dimensions, added, and passed to the next layer. With the decoder, 4x convolution is replaced with transposed convolution, and pooling replaced with linear upsampling. In the case of the conditional model (components with dotted lines) $\mathbf{M_T}$, the reference input $\boldsymbol{x}_r$ is linearly interpolated to be the same size as the output, and passed through a 1x kernel. The target label is passed through a 1x kernel, and added to each pixel of the output. Spectral weight normalization [35] is utilized at every convolutional or fully-connected operation. In the case of the 3D model the convolutions are three-dimensional, and the 2D model uses two-dimensional convolutions.

consists of an encoder that computes the variational posterior, $q(\boldsymbol{z}_r|\boldsymbol{x}_r)$ constructed by serial residual blocks (see Fig. 7) that perform convolutional operations, spatially downsampling the image by half and increasing channel dimension at each layer. The output is then reshaped to a vector and passed independently through two fully connected layers to result in $\boldsymbol{z}_r = N(\mu_{\boldsymbol{z}_r}, \sigma_{\boldsymbol{z}_r})$. $\boldsymbol{z}_r$ is sampled from that normal distribution and passed through a fully connected layer, and passed through residual blocks that spatially upsample, and decrease channel dimension, progressively decoding the latent representation. The same architecture is used for the target model, $\mathbf{M_T}$, but instead the target label and a linearly downsampled copy of $\boldsymbol{x}_r$ is passed in as well.

The primary layer component of this model is a modified residual layer [36], and a detailed description can be found in Fig. 7. Table 3 shows the high-level model architecture.

The 2D model was implemented the same as above but with 2D convolution operations.

The model is trained to maximize the Evidence Lower Bound (ELBO) given an input image $\boldsymbol{x}_{r,t}$:

$$\log p(\boldsymbol{x}_{r,t}|t) \geq \mathrm{ELBO}(\boldsymbol{x}_{r,t}|t) = \mathbb{E}_{q(\boldsymbol{z}_{r,t}|\boldsymbol{x}_{r,t})}[\log p(\boldsymbol{x}_{r,t}|\boldsymbol{z}_{r,t}, t)] - \mathrm{KL}(q(\boldsymbol{z}_{r,t}|\boldsymbol{x}_{r,t}, t)|p(z)) \tag{1}$$

An interpretation of this procedure is that we seek to find a model such that the observed data is the most probable under the model distribution, with the ELBO is as (tractable) approximation of this quantity. Under the generative model described in Fig. 1b, we factor out structure and reference components and train two separate components:

$$\log p(\boldsymbol{x}_{r,t}|t) = \log p(\boldsymbol{x}_t|\boldsymbol{x}_r, t) + \log p(\boldsymbol{x}_r) \geq \mathrm{ELBO}(\boldsymbol{x}_t|\boldsymbol{x}_r, t) + \mathrm{ELBO}(\boldsymbol{x}_r) \tag{2}$$

The embeddings produced by the encoder $q(z|x)$ are encouraged to be compact in the sense that they are penalized for differing in distribution from a standard normal distribution (as measured by $\mathrm{KL}(q(z|x)|q(z))$). An embedding, however compact, is only useful insofar as it is able to faithfully recapitulate the data. The decoder $p(x|z)$ is optimized to produce faithful reconstructions via the reconstruction error term $\mathbb{E}_{q(\boldsymbol{z}_{r,t}|\boldsymbol{x}_{r,t})}[\log p(\boldsymbol{x}_{r,t}|\boldsymbol{z}_{r,t})]$, which encourages the model to balance compactness against transmitting enough information to accurately reconstruct the data.

For section 4.3 we adopt the following reparameterized ELBO definition:

$$\mathrm{ELBO}(x) = (1-\beta)\mathbb{E}_{q(z|x)}[\log p(x|z)] - \beta KL(q(z|x)|p(z)). \tag{3}$$

where $\beta$ is between 0 and 1.

This slight modification to the objective function allowed us to trade-off the relative importance between the reconstruction and sparsity terms of our loss function while keeping the order of magnitude of the total objective function constant. This is greatly beneficial in training multiple models at different values of $\beta$, without needing to modify any other hyper parameters to compensate for an objective function that grows with $\beta$, as in the traditional parameterization of the $\beta VAE$ objective function.

## 6.3 Calculation of Evidence Lower Bound

To calculate the ELBO we use the low-variance estimator in [21]. We use pixel-wise mean squared error to approximate the reconstruction likelihood and average over ten samples from $\boldsymbol{z}_r$ or $\boldsymbol{z}_t$.

## 6.4 Training Procedure

Each model is trained with a batch size of 32 at a learning rate of 0.0002 for 300 epochs via gradient-descent with the Adam optimizer [37]. The optimizer $\beta$ hyper-parameter values are set to (0.9, 0.999) (not to be confused with $\beta$ in the model's objective function). The maximum latent space dimensionality for the reference structures, $\boldsymbol{Z}^r$, and target structures, $\boldsymbol{Z}^s$, were each set to 512 dimensions. We adopt the stochastic training procedure outlined in [21].

We split the data set into 80% training, 10% validation and 10% test, and trained both the reference and conditional model for 300 epochs, and for each training session use the model with the highest ELBO on the validation set.

The model was implemented in PyTorch version 1.2.0, and each component ($P(x_r)$ and $P(x_t|x_r,t)$) was trained on an NVIDIA V100 graphics card. 3D models took approximately two weeks to train while 2D models took approximately 1.5 days to train.

To address overfitting, we evaluate the ELBO on images assigned to the validation set at every epoch. For all results in this manuscript, the model with the highest validation-set ELBO is used. For sparsity/reconstruction models in Fig. 3, we use the unweighted ELBO.

## 6.5 Statistic of Subcellular Structure Coupling

The per-channel statistic we display in Fig. 5b is computed for each cell $\boldsymbol{x}_i$ by considering the likelihood of that cell under the model, compared to the likelihood of that cell with one of its channels swapped out for that same channel from a different cell:

$$c_i^{rs} = \frac{\mathrm{ELBO}(\boldsymbol{x}_i = [\boldsymbol{x}_i^r, \boldsymbol{x}_i^{r'}, \boldsymbol{x}_i^s])}{\frac{1}{N_s}\sum_{j=1}^{N_s}\mathrm{ELBO}(\boldsymbol{x}_{ij} = [\boldsymbol{x}_j^r, \boldsymbol{x}_i^{r'}, \boldsymbol{x}_i^s])} \tag{4}$$

Here $r$ is the reference channel (either the membrane or the nucleus) that we are evaluating, $s$ is the structure type, denoting which set of cells we aggregate over. $\boldsymbol{x}_i = [\boldsymbol{x}_i^r, \boldsymbol{x}_i^{r'}, \boldsymbol{x}_i^s]$ is the three channel image decomposed into the reference channel of interest $r$, the other reference channel $r'$, and the structure channel $s$. The numerator is the likelihood of the original data, and the denominator is the average permuted likelihood of the cell with the reference channel of interest $r$ permuted across all other cells with the same structure tagged (i.e. $N_s$ cells).

To aggregate this per-channel coupling strength into a *relative* coupling value, we take the ratio of the difference over the sum of the membrane-structure coupling and the nucleus-structure coupling. That is, the differential coupling of a structure to the membrane vs the nucleus $d_s$ is computed as

$$d_s^p = \frac{1}{N_s^p} \sum_{i=1}^{N_s^p} \frac{c_i^{ms} - c_i^{ns}}{c_i^{ms} + c_i^{ns}} \tag{5}$$

where $N_s^p$ is the number of cells where structure $s$ is tagged and are also in cell cycle phase $p$ (interphase or mitosis), $c_i^{ms}$ is the coupling of structure $s$ in cell $i$ to the membrane, and similarly $c_i^{ns}$ is the coupling of the structure in that cell to the nucleus.

# 7 Acknowledgments

We thank Rick Horwitz for scientific discussions. We thank Neda Bagheri and Irina Mueller for careful reading and commenting on the manuscript. We thank Graham Johnson and Thao Do for help with visualization. We thank Roy Wollman for valuable advice. We thank the Allen Institute for Cell Science team for creation of the data that enabled this study and for helpful scientific discussions and support. We wish to acknowledge the Allen Institute for Cell Science founder, Paul G. Allen, for his vision, encouragement and support.

# 8 Supporting information

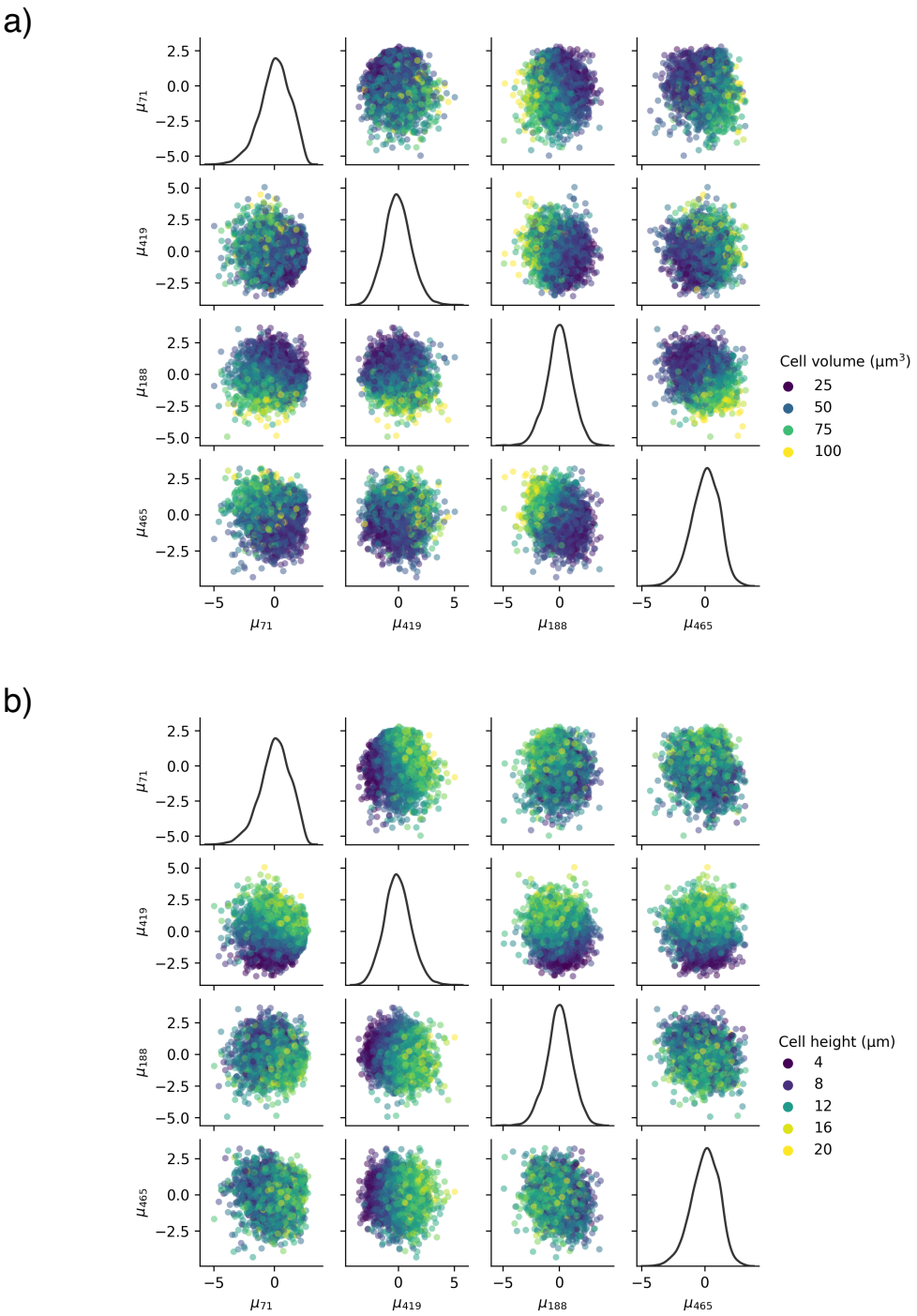



**Fig S1.** Pairwise plots of the top four latent space dimensions, as ranked by mean absolute deviation from 0 on the test set. The marginal distribution of each latent dimension is plotted on the diagonal. a) Here we color by the cell volume, and see a visually apparent pattern in the data — most notably a strong correlation with $\mu_{188}$. b) Here we color by the cell height, and again observe structure in the scatter plots — most notably a strong correlation with $\mu_{419}$.

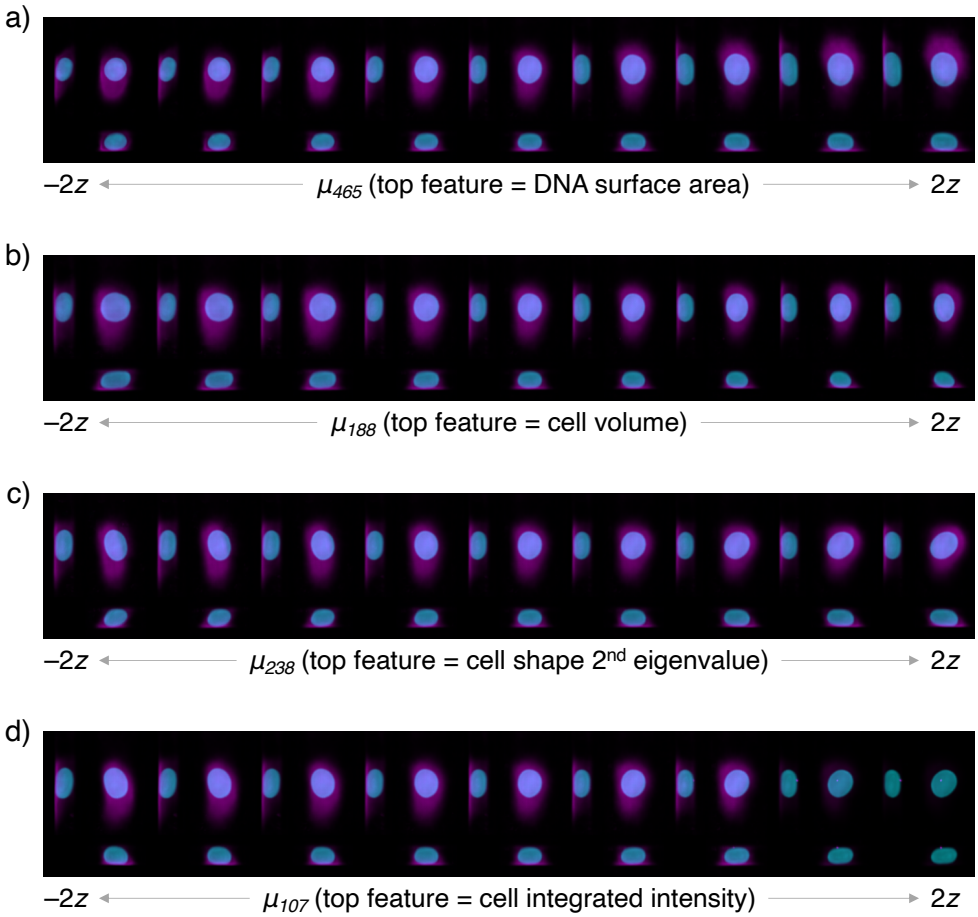



**Fig S2.** Latent space walks along the 3rd through 6th highest ranked dimensions, as ranked by mean absolute deviation from 0 on the test set. Walks are performed along the specified dimension in nine steps, starting at negative two standard deviations and ending at two standard deviations. All other latent dimensions are set to 0. We include the name of the most highly correlated cell feature, but the cell features are highly correlated (see Supp. Fig. S3) and a single latent space dimension may correlate with many cell features. a) Latent dimension $\mu_{465}$, which is most strongly correlated with nuclear surface area. b) Latent dimension $\mu_{188}$, which is most strongly correlated with cell volume. c) Latent dimension $\mu_{238}$, which is most strongly correlated with tilt/shear along the $x$-$z$-axes. d) Latent dimension $\mu_{107}$, which is most strongly correlated with the total integrated intensity in the plasma membrane dye channel.

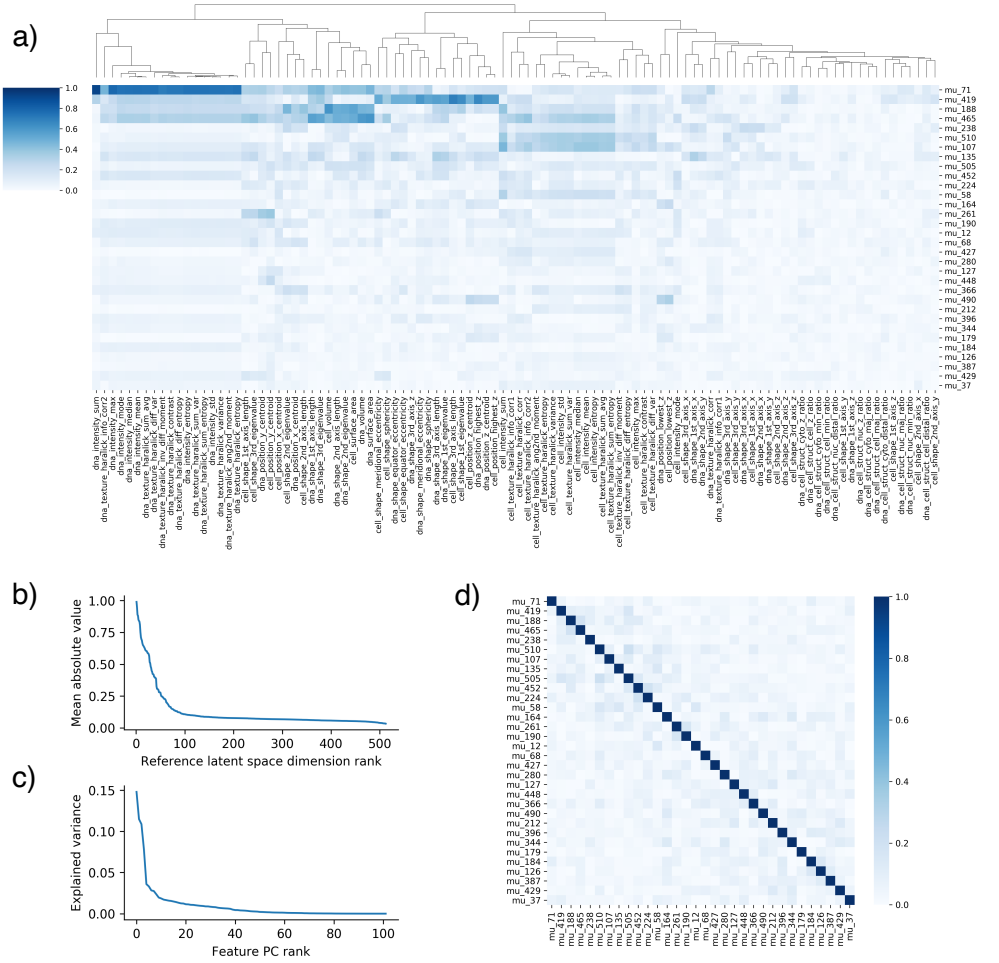



**Fig S3.** a) Heatmap of Spearman correlations of reference latent space dimensions with single-cell features on all cells in the test set. Cell features are hierarchically clustered. Latent space dimensions are sorted in descending rank by mean absolute deviation from 0, and for clarity only the top 32 dimensions are shown. Dimensions below 32 displayed significantly more noise and less correlation with cell features. b) Mean absolute deviation from 0 of all reference latent space dimensions, sorted by value. Values are computed by averaging over all cells in the test set. c) Explained variance of principal components of the z-scored cell features on all cells in the test set. d) Pearson correlation of the top 32 dimensions of the latent space, computed on all cells in the test set as ranked by mean absolute deviation from 0. We note that these dimensions display little to no correlation structure, empirically verifying the ability of the $\beta$-VAE to produce a disentangled latent space.

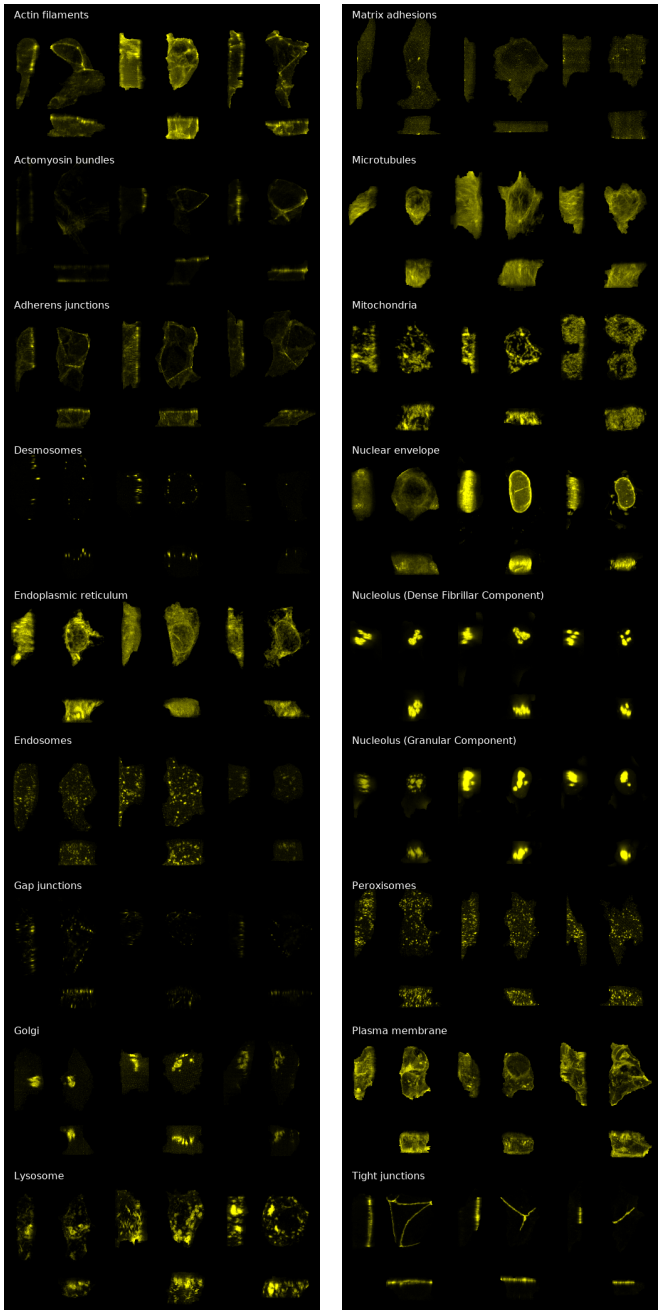



**Fig S4.** Three examples of each mEGFP-tagged structure are shown, sampled randomly from our test data set. Each cell only has one mEGFP-tagged structure, so examples are all from different cells.

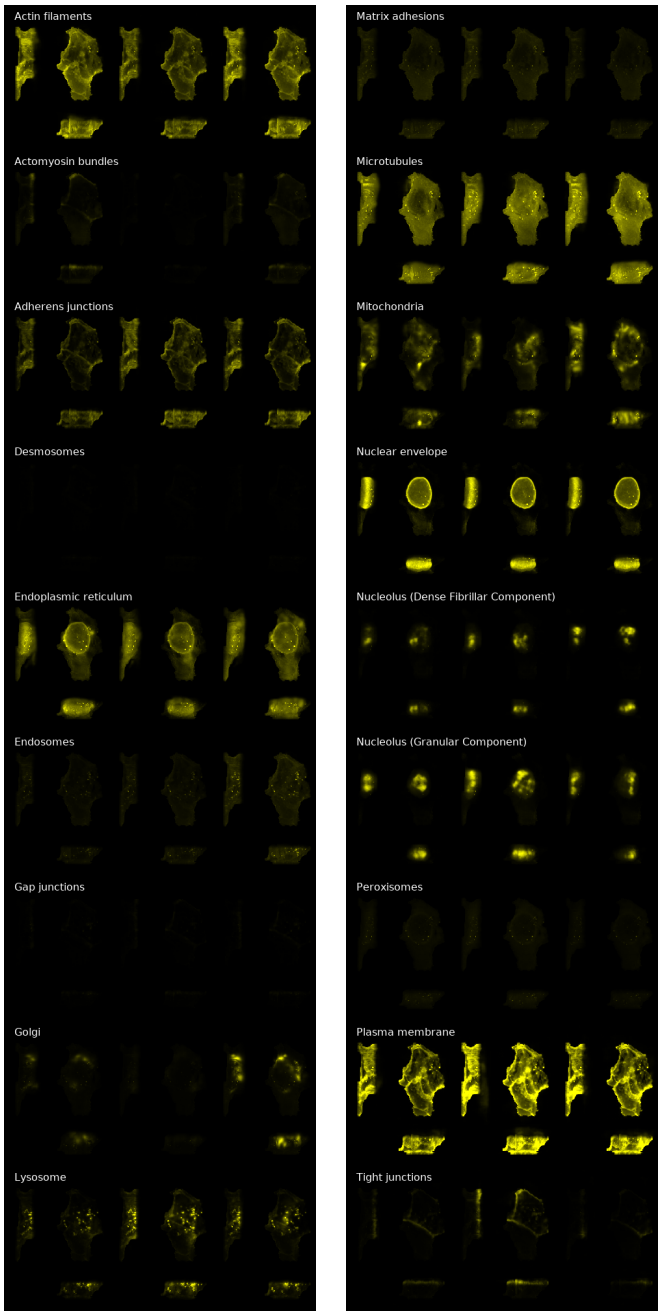



**Fig S5.** Structures generated by our model. Three examples of each mEGFP-tagged structure are shown. Structures are generated using random draws from the conditional latent space, while keeping the reference geometry fixed to a single (randomly chosen) cell geometry from the test set. The same cell geometry is used across all structures shown here.

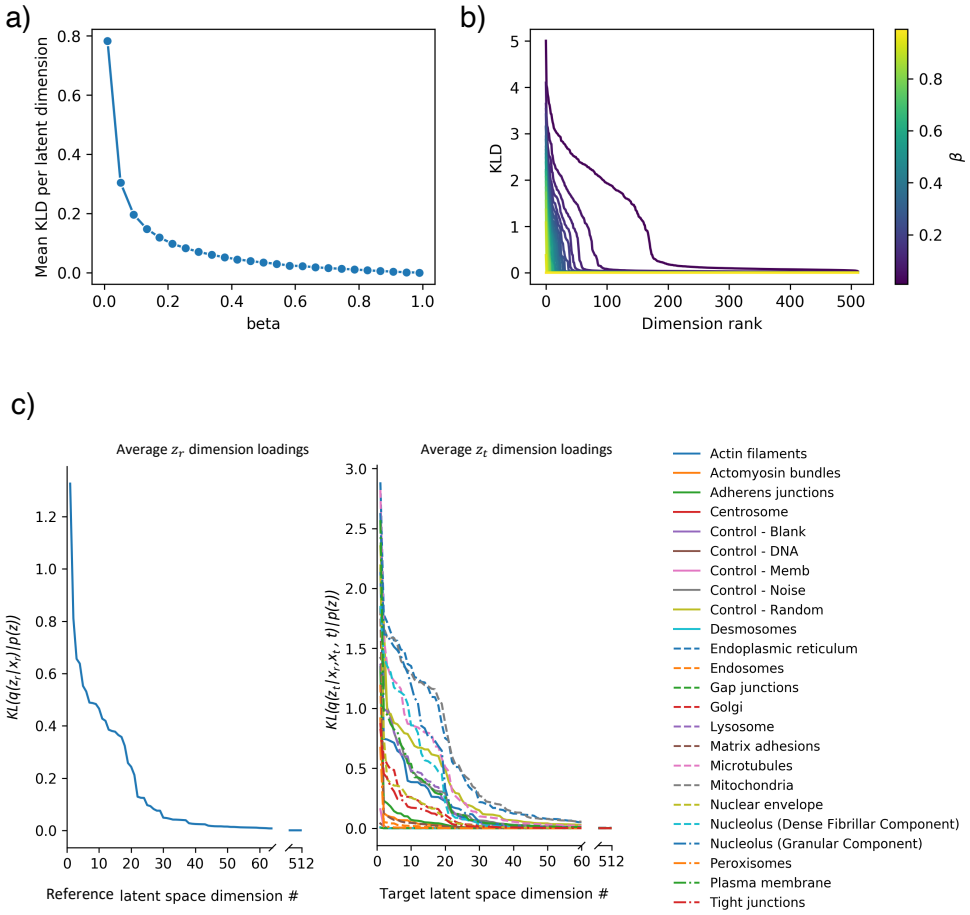



**Fig S6.** a) Mean KLD per dimension for the reference latent space of the test set in the 2D model, as a function of $\beta$, averaged over all dimensions in the latent space. b) Mean KLD per dimension for the reference latent space of the test set in the 2D model, as a function of dimension rank, for each model fit using a different $\beta$. c) Left: Mean KLD per dimension for the reference latent space of the test set in the 3D model, as a function of dimension rank. Right: Mean KLD per dimension for the conditional latent space of the test set in the 3D model, as a function of dimension rank and structure type.

# References

1. Murphy RF. Location proteomics: a systems approach to subcellular location. Biochemical Society transactions. 2005;33(Pt 3):535–538.

2. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509(7502):575–581.

3. Zhao T, Murphy RF. Automated learning of generative models for subcellular location: Building blocks for systems biology. Cytometry Part A. 2007;71A(12):978–990.

4. Peng T, Murphy RF. Image-derived, three-dimensional generative models of cellular organization. Cytometry Part A. 2011;79A(5):383–391.

5. Majarian TD, Cao-Berg I, Ruan X, Murphy RF. CellOrganizer: Learning and Using Cell Geometries for Spatial Cell Simulations. In: Modeling Biomolecular Site Dynamics. Springer; 2019. p. 251–264.

6. Donovan RM, Tapia JJ, Sullivan DP, Faeder JR, Murphy RF, Dittrich M, et al. Unbiased rare event sampling in spatial stochastic systems biology models using a weighted ensemble of trajectories. PLoS computational biology. 2016;12(2):e1004611.

7. Vasan R, Rowan MP, Lee CT, Johnson GR, Rangamani P, Holst M. Applications and Challenges of Machine Learning to Enable Realistic Cellular Simulations. Frontiers in Physics. 2020;7:247.

8. Boland MV, Murphy RF. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics. 2001;17(12):1213–1223.

9. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology. 2006;7(10):R100.

10. Rajaram S, Pavie B, Wu LF, Altschuler SJ. PhenoRipper: software for rapidly profiling microscopy images. Nature Methods. 2012;9(7):635–637.

11. Chen J, Ding L, Viana MP, Hendershott MC, Yang R, Mueller IA, et al. The Allen Cell Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. bioRxiv. 2018; p. 491035.

12. Van Valen DA, Kudo T, Lane KM, Macklin DN, Quach NT, DeFelice MM, et al. Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. PLOS Computational Biology. 2016;12(11):e1005177.

13. Ounkomol C, Seshamani S, Maleckar MM, Collman F, Johnson GR. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Nature methods. 2018;15(11):917.

14. Christiansen EM, Yang SJ, Ando DM, Javaherian A, Skibinski G, Lipnick S, et al. In silico labeling: predicting fluorescent labels in unlabeled images. Cell. 2018;173(3):792–803.

15. Weigert M, Schmidt U, Boothe T, Müller A, Dibrov A, Jain A, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. Nature methods. 2018;15(12):1090–1097.

16. Ouyang W, Aristov A, Lelek M, Hao X, Zimmer C. Deep learning massively accelerates super-resolution localization microscopy. Nature biotechnology. 2018;36(5):460–468.

17. Waisman A, La Greca A, Möbbs AM, Scarafía MA, Velazque NLS, Neiman G, et al. Deep learning neural networks highly predict very early onset of pluripotent stem cell differentiation. Stem cell reports. 2019;12(4):845–859.

18. Eulenberg P, Köhler N, Blasi T, Filby A, Carpenter AE, Rees P, et al. Reconstructing cell cycle and disease progression using deep learning. Nature communications. 2017;8(1):1–6.

19. Kimmel JC, Brack AS, Marshall WF. Deep convolutional and recurrent neural networks for cell motility discrimination and prediction. bioRxiv. 2019; p. 159202.

20. Lu AX, Kraus OZ, Cooper S, Moses AM. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. PLoS computational biology. 2019;15(9):e1007348.

21. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013;.

22. Alemi AA, Poole B, Fischer I, Dillon JV, Saurous RA, Murphy K. Fixing a broken ELBO. arXiv preprint arXiv:171100464. 2017;.

23. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. Iclr. 2017;2(5):6.

24. Sønderby CK, Raiko T, Maaløe L, Sønderby SK, Winther O. Ladder variational autoencoders. In: Advances in neural information processing systems; 2016. p. 3738–3746.

25. Johnson GR, Donovan-Maiye RM, Maleckar MM. Generative Modeling with Conditional Autoencoders: Building an Integrated Cell. arXiv preprint arXiv:170500092. 2017;.

26. Johnson GR, Donovan-Maiye RM, Maleckar MM. Building a 3D Integrated Cell. bioRxiv. 2017;doi:10.1101/238378.

27. Goldsborough P, Pawlowski N, Caicedo JC, Singh S, Carpenter A. CytoGAN: Generative Modeling of Cell Images. bioRxiv. 2017; p. 227645.

28. Osokin A, Chessel A, Carazo Salas RE, Vaggi F. GANs for biological image synthesis. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2233–2242.

29. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:171010196. 2017;.

30. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 8798–8807.

31. Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:170100160. 2016;.

32. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial Autoencoders. arXivorg. 2015;.

33. Hou X, Shen L, Sun K, Qiu G. Deep Feature Consistent Variational Autoencoder. CoRR. 2016;abs/1610.00291.

34. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature methods. 2018;15(12):1053.

35. Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:180205957. 2018;.

36. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

37. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.