Virtual perturbations to assess explainability of deep-learning based cell fate predictors

Christopher J. Soelistyo
University College London, UK
Alan Turing Institute, UK
csoelistyo@turing.ac.uk

Guillaume Charras
University College London, UK
g.charras@ucl.ac.uk

Alan R. Lowe
University College London, UK
Alan Turing Institute, UK
alowe@turing.ac.uk

Abstract

Explainable deep learning holds significant promise in extracting scientific insights from experimental observations. This is especially so in the field of bio-imaging, where the raw data is often voluminous, yet extremely variable and difficult to study. However, one persistent challenge in deep learning assisted scientific discovery is that the workings of artificial neural networks are often difficult to interpret. Here we present a simple technique for investigating the behavior of trained neural networks: virtual perturbation. By making precise and systematic alterations to input data or internal representations thereof, we are able to discover causal relationships in the outputs of a deep learning model, and by extension, in the underlying phenomenon itself. As an exemplar, we use a recently described deep-learning based cell fate prediction model. We trained the network to predict the fate of less fit cells in an experimental model of mechanical cell competition. By applying virtual perturbation to the trained network, we discover causal relationships between a cell's environment and eventual fate. We compare these with known properties of the biological system under investigation to demonstrate that the model faithfully captures insights previously established by experimental research.

1. Introduction

1.1. Deep learning and scientific discovery

Scientific models are typically built to explain events in the natural world. This often consists of building relationships between elements of the world; for example, the relationship between the current environment of a biological cell and its eventual mitotic or apoptotic fate.

The discovery of patterns within observed data is a key strength of deep learning. Therefore, its recent emergence has raised the potential for the automated generation of accurate scientific models [7]. A deep neural network (DNN), through being trained on large amounts of real-world data, can create internal representations that capture relationships in the natural world - in other words, a scientific model that can predict the outcome of experimental observations. This stems from the incredible power of DNNs to approximate input-output mappings given a sufficient corpus of training examples.

Abstractly, the aim of machine learning (ML) in general is to produce a model that can approximate some desired mapping f - which represents the natural phenomenon - from an input domain X to an output domain Y:

$$f : X \to Y \quad X, Y \subset \mathbb{R}^n. \tag{1}$$

The goal is to learn some function g that approximates f (i.e., $g(x) \approx f(x)$). This approximate function could be modelled by a DNN.

1.2. The problem of scientific explainability

There are at least two key issues with using DL to generate scientific models. The first is that it is notoriously difficult to explain how a deep neural network arrives at its outputs. This is due largely to the typical complexity of DNNs, rendering them "black boxes" resistant to human interpretation. Therefore, it can be difficult to build accurate DNNs that are simultaneously explainable. However, we hypothesize that it is possible to extract insights learned by the DNN by focusing on causal relationships between

input and output, a theme that will be further developed in Section 3 within the context of deep-learning based cell fate predictors.

The other key issue with DL-based science is that it may be possible to develop a multitude of models that achieve the same degree of accuracy through a diversity of internal mechanisms. Hence the learned function, or "theory", g could belong to a wider set G of equal-performance functions, what Semenova et. al [26] call a "Rashomon set". Therefore, by fitting the data, a trained model will not necessarily have learned the underlying natural phenomenon; it will have learned *a* theory, perhaps one of many, that is consistent with the observed data. We will return to this point in the discussion (Section 4).

1.3. Deep learning as a tool in bioimaging

One of the scientific fields most promising for the application of deep learning is imaging of cell biological processes. This is due primarily to the inherent complexity and volume of biological data, as well as its tractability as a problem. The continued development of microscopy methods has made available a wealth of image data related to cell appearance, organization and behavior. Meanwhile, automation and cell segmentation technology has further enabled the high-throughput collection of vast amounts of spatio-temporal data, allowing for the capture of time-lapse videos [36].

This abundance of data provides both opportunities and challenges for the cell imaging field. Cell biologists can access an unprecedented amount of information. However, as noted by Ouyang and Zimmer [22], the volume of data produced by modern imaging technologies has "outgrow[n], often vastly, the capacity of manual analyses and human inspection". As a result, this field has seen a recent explosion of studies applying deep learning [5]. The capability of DNNs to extract patterns from complex data has led to their application by cell biologists in tasks as varied as feature extraction [13, 23, 29, 31], morphology-based classification [11, 24, 33, 14, 37, 18, 39], image segmentation [25, 6, 30, 10, 34, 1, 27, 4] , synthetic data generation [21, 8] and more.

In recent years, the field has also seen applications of explainable deep learning to extract scientifically relevant patterns from complex bioimaging data, for example to identify indicators of metastatic efficiency in melanoma cells [38] or determinants of cell fate in cell competition [28]. The present study aims to extend this latter work, by systematically examining input-output relationships in a DNN model designed to predict cell fate.

2. Model system: predicting cell fate using a deep neural network

Our model system is a competition between wild-type MDCK cells (MDCK^{WT}) and scribble-knockdown MDCK cells (scrib^{kd}), where the latter expresses an shRNA that silences expression of the Scrib gene. The loss of the scribble polarity protein is a deleterious mutation, and triggers competitive interactions between the MDCK^{WT} and scrib^{kd} cells, where the presence of the former leads to elimination of the latter via apoptosis.

A decade of experimental research has revealed the dynamics of MDCK^{WT} and scrib^{kd} cell competition [19, 35, 2, 9]. MDCK^{WT} cells have higher homeostatic density than scrib^{kd} cells, therefore, when the populations are put in contact, the presence of MDCK^{WT} cells induces crowding in scrib^{kd} cells, which leads to their mechanical compression and eventual elimination via apoptosis [35]. The key role of crowding in fate determination in scrib^{kd} cells was then independently rediscovered by a deep neural network trained to predict cell fate (apoptosis vs. mitosis) based on time-lapse videos of single cells taken by fluorescence microscopy [28].

In this latter study, Soelistyo et. al [28] build a three-stage neural network architecture (referred to as the τ-VAE) to predict cell fate from a sequence of images. In particular, the network converts a time-lapse video portraying a scrib^{kd} cell throughout its lifetime, called a "trajectory", into a prediction of the cell's eventual fate. These trajectories were obtained by using the *btrack* package [17, 34] on U-Net [25] segmentations of cell nuclei. In each video, MDCK^{WT} and scrib^{kd} nuclei are distinguished because their Histone 2B protein is tagged with different fluorophores, green fluorescent protein (GFP) for MDCK^{WT} and red fluorescent protein (RFP) for scrib^{kd}. Importantly, the videos do not contain any evident morphological changes of the nucleus prior to the fate event. This ensures that the model is trained to *predict* cell fate, rather than simply *observe* morphological indicators of cell fate.

The three-stage τ-VAE consists of the encoder part of a variational autoencoder [15, 12, 3], a principal component (PC) projective transform and a temporal convolutional network (TCN) predictor [20] (**Fig. 1**). The PC-transform step was necessary to obtain embedded representations that were interpretable. The results were most striking for PC0 and PC1, which represented cell type and cell size/nuclear area respectively.

The TCN was trained to classify scrib^{kd} cell trajectories into one of two potential fates: mitosis and apoptosis. A third "synthetic" class was added to represent those inputs that fall outside of the distribution (OOD) of real cell trajectories. Data for this class was artificially generated using a random-walk procedure in latent space [28]. The "syn-

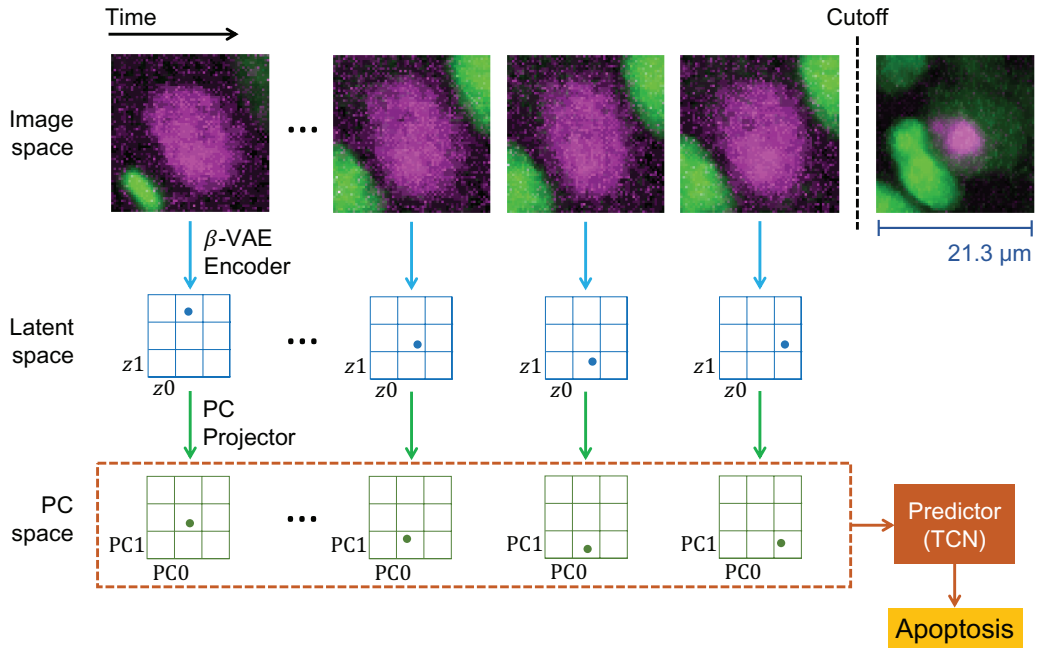

Figure 1. **τ-VAE pipeline, from trajectory to prediction.** The input is the trajectory in image space, trimmed to a cutoff such that it does not contain the morphological features that identify a fate event. Images are transformed into latent space by the β-VAE encoder, then into PC space by the PC projector, where they are collectively input to the TCN predictor. In the current example, a cell is undergoing apoptosis in response to crowding.

thetic" class was added to ensure that the τ-VAE learned the predictors of both mitosis and apoptosis, rather than learning the predictors of one and "dumping" all remaining trajectories in the other fate class.

The PCs most important to cell fate prediction were then identified using an ablation procedure. This confirmed that the τ-VAE had learned that nuclear area was the most important determinant of cell fate, completely independently of all the scientific studies that had arrived at the same conclusion.

The aim of the present study is to extend this work by characterizing the impact of certain input features on the predictive behavior of the τ-VAE. In particular, we sought to characterize the sensitivity of the model's predictions to perturbations in both its internal representation of the input data, and perturbations in the input data itself.

To ensure continuity with the work of Soelistyo et. al [28], we followed the same procedure for data acquisition, image processing, model training and synthetic trajectory generation. We also implemented PCA in the equivalent manner, using the PCA class from Scikit-learn. The projective transformation that maps points in latent space to PC space consisted of the operation:

$$X_{PC} = (X_{latent} - \mu) \cdot C^{\intercal}, \tag{2}$$

where X_{PC} and X_{latent} represent corresponding positional vectors in PC space and latent space, μ is a vector of per-component mean values, and C is a matrix of weights that define how each PC is calculated from each latent variable. The inverse projection is:

$$X_{latent} = (X_{PC} \cdot C) + \mu. \tag{3}$$

3. Model perturbations

Perturbation is commonly used to uncover relationships between cause and effect. By varying one independent variable while keeping others constant, the specific contribution of that variable can be measured in a way that minimizes exposure to confounding factors. This minimization can be achieved by recording the statistically significant effects of a perturbation throughout a highly diverse population of inputs.

This logic allows us to examine the predictive behavior of a deep neural network (or any other input-output model, in fact). By applying specific perturbations to the data representation at some point in the information processing pipeline, we can uncover causal relationships between model inputs and outputs. This can be either a physical perturbation, such as treating the cell culture with some drug, or a *virtual* perturbation, which involves altering the model's internal representations of the underlying system without affecting the system itself (**Fig. 2**).

When applying a perturbation to the internal representation, it is often important that this representation is disentan-

gled to a degree that enables the identification of particular latent features with specific physical/conceptual attributes, such as cell type or cell size. While achieving perfect disentanglement is extremely difficult, our β-VAE managed this to a high degree.

More generally, we can conceptualize a perturbation as some intervention applied to the data representation at some location in the pipeline (**Fig. 2**). In our case, the pipeline can be expressed as:

$$\begin{aligned} f:\ X &\to Y \\ g:\ X &\to \hat{Y} \\ g(X) &\approx f(X) \\ g(X) &= T(R(E(M(X)))) \\ \hat{Y} &= T(R(E(M(X)))) \end{aligned} \tag{4}$$

where X is the system under study (i.e., the cell competition), M is the imaging and tracking procedure, E is the β-VAE encoder, R is the PC projection and T is the TCN predictor. Y is then the true cell fate and $\hat{Y}$ is the predicted cell fate.

In this study, we focus on perturbations (denoted by Δ) applied to the time-lapse input data, post-tracking (Eq. 5), as well as on the model's final internal representation of the input, in PC space (Eq. 6).

$$\hat{Y} = T(R(E(\Delta + M(X)))). \tag{5}$$

$$\hat{Y} = T(\Delta + R(E(M(X)))). \tag{6}$$

We applied perturbations that would allow us to further test the findings of our previous model and also represent situations that could plausibly occur in the underlying system, and so lie within the distribution of input trajectories used to train the τ-VAE (hereafter called the "realistic range"). So, for example, switching the cell type of neighboring cells is a valid operation because it produces a situation that could plausibly occur in this range. Meanwhile, switching the cell type of the central cell is not valid because the τ-VAE was trained on scrib^{kd} cells and not MDCK^{WT} cells, hence switching the central cell type to MDCK^{WT} would push the perturbed trajectories outside the realistic range.

For the manipulation of input data, we chose neighbor cell type switching because it allowed us to test the conclusion of our own and other studies that neighbor cell type is not an important determinant of cell fate in scrib^{kd} cells [35, 2, 28]. Similarly, for the manipulation of internal representations, we decided to adjust PC1 because it allowed us to the test the finding that crowding/nuclear compression *is* an important determinant.

3.1. Manipulation of the input data

Wagstaff et. al [35] demonstrated that the presence of MDCK^{WT} cells induces apoptosis in scrib^{kd} cells only indirectly, by promoting greater crowding and a higher local cell density, which in turn mechanically compresses the scrib^{kd} cells. As the scrib^{kd} loser cells are less tolerant to crowding, they commit apoptosis. This suggests that the actual identity of the neighbors is not important, rather it is the degree of crowding/compression alone that determines the fate.

Therefore, switching the cell type of neighboring cells should have a small effect on cell fate prediction by the β-VAE, given that the perturbed trajectory should show similar degrees of compression to the original trajectory. To switch the cell type of neighbors, the U-Net segmentation mask was first used to determine the locations of the cells in an image. Then, in these locations, the maximum pixel value across both channels (GFP and RFP) was placed in the channel corresponding to the cell type opposite to the original cell type. We spared the central cell region from this operation. The whole operation can then be described by:

$$I \to (M \odot I_{switched}) + (M_{inverse} \odot I), \tag{7}$$

where I is the image, M is the binary mask denoting cell pixels, $I_{switched}$ is a channel-switched representation of I, and $M_{inverse}$ is the negation of M. The $\odot$ operator is the Hadamard (element-wise) product.

Due to inherent intensity differences between the GFP and RFP signals, these were normalized throughout the original image on a per-channel basis prior to switching. We normalized each channel by first applying a median filter then removing outliers (those pixels whose values deviated from the local median by some threshold). Then, we calculated the 5^{th} and 99^{th} percentile pixel values, divided the image by the difference between these, and subtracted the 5^{th} percentile value from the result. Finally, we clipped the image to values between 0.0 and 1.0. After this step, the images are input through the three-stage τ-VAE as usual (**Fig. 3**).

The predictive results of the τ-VAE on the perturbed dataset are shown in **Fig. 4**. Results were obtained by testing ten models on ten different testing sets (with $\sim$300 scrib^{kd} cell trajectories each), obtained by 10-fold cross validation. Mean and standard deviation across the ten models were used to produce the confusion matrices (**Fig. 5**). The variation, compared to the unperturbed dataset, was very slight. There were small differences; in particular, the rate of mitosis prediction seemed to have increased marginally. This could indicate that the τ-VAE uses information about neighbor cell type to a small degree. However, it could also reflect the fact that when the cell type of neighbor cells is switched using our method, the τ-VAE's representations of the *central* cell are also affected (as shown in **Fig. 3**). This could in turn reflect the fact that certain configurations of cells are found only sparsely in the image dataset on which the β-VAE was trained. In particular, situations where a scrib^{kd} cell is surrounded by other scrib^{kd} cells are quite

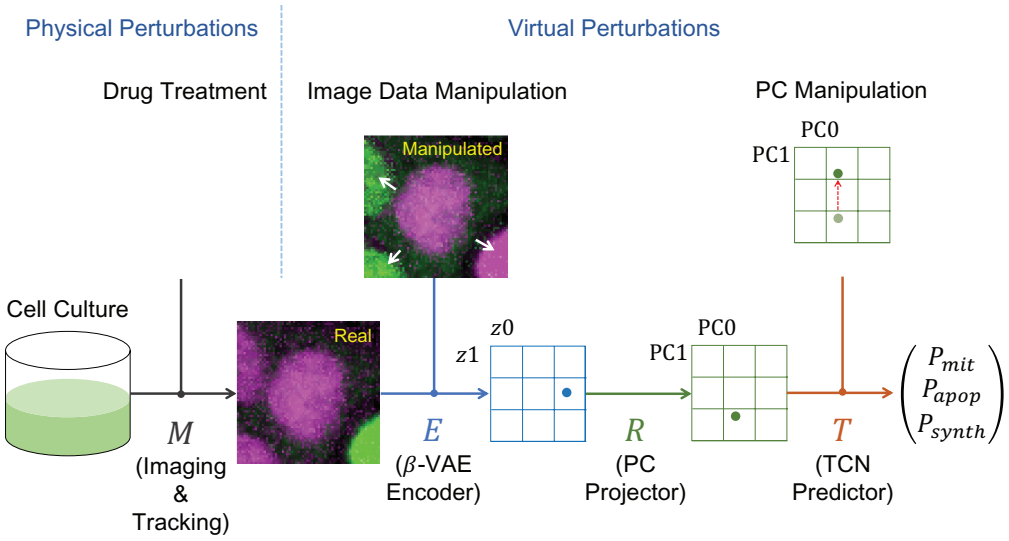

Figure 2. **Virtual perturbation schemes.** Location of the perturbations within the data pipeline. neighbor cell type switching occurs before the encoding step (white arrows show perturbed cells). Nuclear area/PC1 adjustment occurs after PC projection. Drug treatments are applied directly to the cells before imaging (this perturbation does not feature in the present study, but was used by Soelistyo et. al [28]).

rare.

Nevertheless, this perturbation step has demonstrated that the effect on predictor output of switching the cell type of neighbor cells is marginal at best, especially when compared to the large effects of adjusting PC1 - a perturbation whose consequences are of a similar magnitude in image space (Section 3.2).

As a further control, we tested the effects of switching the identity of the central cell (from scrib^{kd} to MDCK^{WT}) rather than the neighbor cells. This led to the τ-VAE assigning a "synthetic" prediction to virtually all ground-truth apoptotic and mitotic trajectories. This result was expected due to the fact that the τ-VAE model under inspection was trained on real trajectories centred on scrib^{kd} cells only. Hence, MDCK^{WT} trajectories are out-of-distribution (OOD) with respect to the data on which the τ-VAE was trained, and the τ-VAE is evidently sensitive to this. OOD effects have implications for the scientific use of deep learning, as will be discussed later.

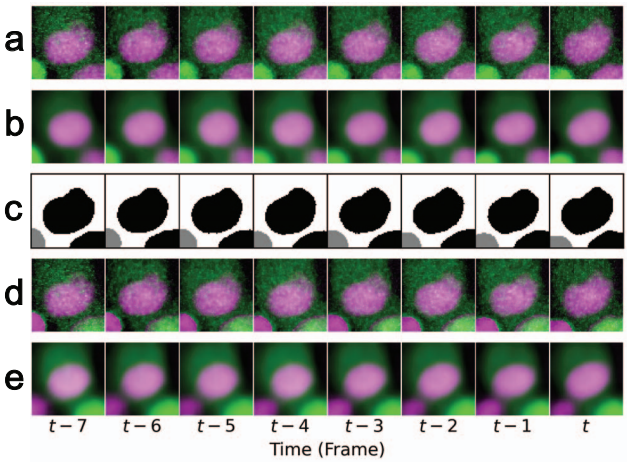

Figure 3. **Neighbor cell-type switching procedure.** Consecutive frames of one trajectory. a) the original raw images, b) the β-VAE reconstructions of the original images, c) the U-Net derived segmentation masks (grey = MDCK^{WT}, black = scrib^{kd}), d) the perturbed images, and e) the β-VAE reconstructions of the perturbed images.

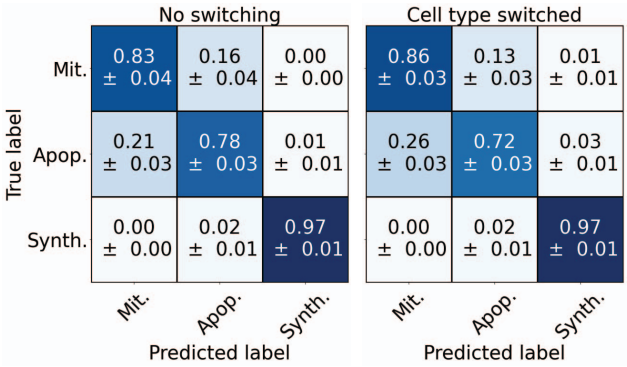

Figure 4. **Confusion matrix showing the performance of the τ-VAE models on testing sets which have been perturbed by neighbor cell type switching.** left) Confusion matrix on the un-perturbed dataset. right) Confusion matrix on the perturbed dataset, demonstrating that the identity of neighboring cells has little effect on the predicted cell fate.

3.2. Manipulation of the internal representation

Mechanical compression of the nucleus, visually represented by a reduction of the nuclear area, has been shown experimentally to induce apoptosis in scrib^{kd} cells, even in the absence of MDCK^{WT} cells [35]. We therefore hypothesized that by virtually decreasing the area of the target cell nuclei in our images, we could increase the false-positive rate of apoptosis detection by our τ-VAE in a predictable manner.

The method we used to adjust nuclear area was to simply increase or decrease PC1, by a constant absolute value, throughout a trajectory (**Fig. 5a,b**). This step occurs after encoding by the β-VAE encoder and projection by the PCA model (Fig. 2). For a given trajectory with T time-points, and a given adjustment value of v, the transformation is:

$$X_{t,1} \rightarrow X_{t,1} + v \quad \text{for} \quad t = 1, 2, ..., T, \tag{8}$$

where X is a matrix of PC-space positional vectors; each row represents one time-point and each column represents one PC.

Since the TCN receives the PC representation of the trajectory, this step could be implemented prior to input to the TCN. Results were obtained by testing ten models on ten different testing sets, in the manner described in Section 3.1. Interestingly, the effect of adjusting PC1 is to simultaneously vary the nuclear area of the central cell, and the degree of crowding in the neighborhood. This reflects the principle, learned by the β-VAE, that higher local density of cells entails a higher degree of nuclear compression.

Virtual perturbations to PC1 induce a predictable, stable effect on the τ-VAE's predictions. Increasing PC1 decreases the rate of mitosis prediction, while decreasing PC1 does the reverse. With increasing PC1, the rate of apoptosis prediction first increases, then decreases as the rate of synthetic prediction becomes dominant. This latter result shows that trajectories with such artificially high values of PC1 in fact do not lie within the distribution of real trajectories; in other words, within the "realistic range". So, for trajectories which *do* lie within the distribution, the picture is one of monotonic decrease in mitotic probability and monotonic increase in apoptotic probability as PC1 is increased.

We have demonstrated that the specific adjustment of PC1 produces predictable and significant effects on the outputs of the τ-VAE. Our hypothesis is that these effects are due to the *nature* of the perturbation (nuclear area adjustment) and not simply its magnitude. To test this, we repeated the same procedure for the other thirty-one PCs aside from PC1.

In general, we did not see the same stable, predictable and monotonic relationship (within the realistic range) as with PC1. A case in point is PC2, whose adjustment induces a small rotation of the major axis of the central cell (**Fig. 5b**). Adjustment of PC2 produces very minor variations in the treatment of ground-truth apoptoses, and it triggers small decreases in performance in both the positive and negative direction for ground-truth mitoses (**Fig. 5d**). Evidently, the adjustment of PC2 is not sufficient to convert mitotic predictions to apoptotic predictions or *vice versa*.

The corresponding confusion matrices are shown in Fig. 5e. To demonstrate the effects of perturbation, we show predictions results for PC1 shifts of different sign. For PC1, we show the results of a $+3.0$ shift rather than a $+6.0$ shift because shifts greater than $+3.0$ tend to the push the trajectories outside of the realistic range, indicated by a large number of synthetic predictions. This is to ensure that we investigate only those trajectories that remain in the distribution of real-world data.

To quantify the relationship between τ-VAE output and each PC, we calculated the gradients of linear models fitted to the prediction curves by ordinary least-squares regression. The intuition is that PCs that monotonically changed the output prediction will show large gradients. In order to capture the PC-output relationship solely within the realistic range, we considered only those data-points for which less than 20% of ground-truth trajectories in a particular class were predicted as synthetic (indicative of the realistic range). As can be seen in Fig. 6, PC1 was associated with the greatest slope magnitudes, demonstrating that its relationship with model output was the most significant and monotonic. This suggests that certain physical features are important for cell fate determination (e.g. PC1/crowding) while some are not (e.g. PC2/rotation).

4. Discussion

Virtual perturbation is a straightforward technique for assessing the impact of various input features on the output of a model, black box or otherwise. In this study we applied this method to a DNN trained to predict cell fate from image sequences, acquired in an experimental system that captures the phenomenon of mechanical cell competition. We show that altering nuclear area in the internal representation of the DNN exerts a significant, predictable and monotonic effect on the rate of apoptosis prediction within the realistic range. In contrast, flipping the cell type of neighboring cells triggers only a minimal effect. Through this, we can confirm that our τ-VAE model has autonomously learned that mechanical compression is a driving factor in apoptosis, while cell-type-dependent interactions with neighbors are not. Thus our deep-learning based approach has recovered the conclusions of previous experimental and statistical research on the system under study [35, 2].

Here, we explore only two types of virtual perturbation to our model system. One could imagine a multitude of others, each testing a different hypothesis. For example, by adjusting PC1 for varying lengths of time, instead

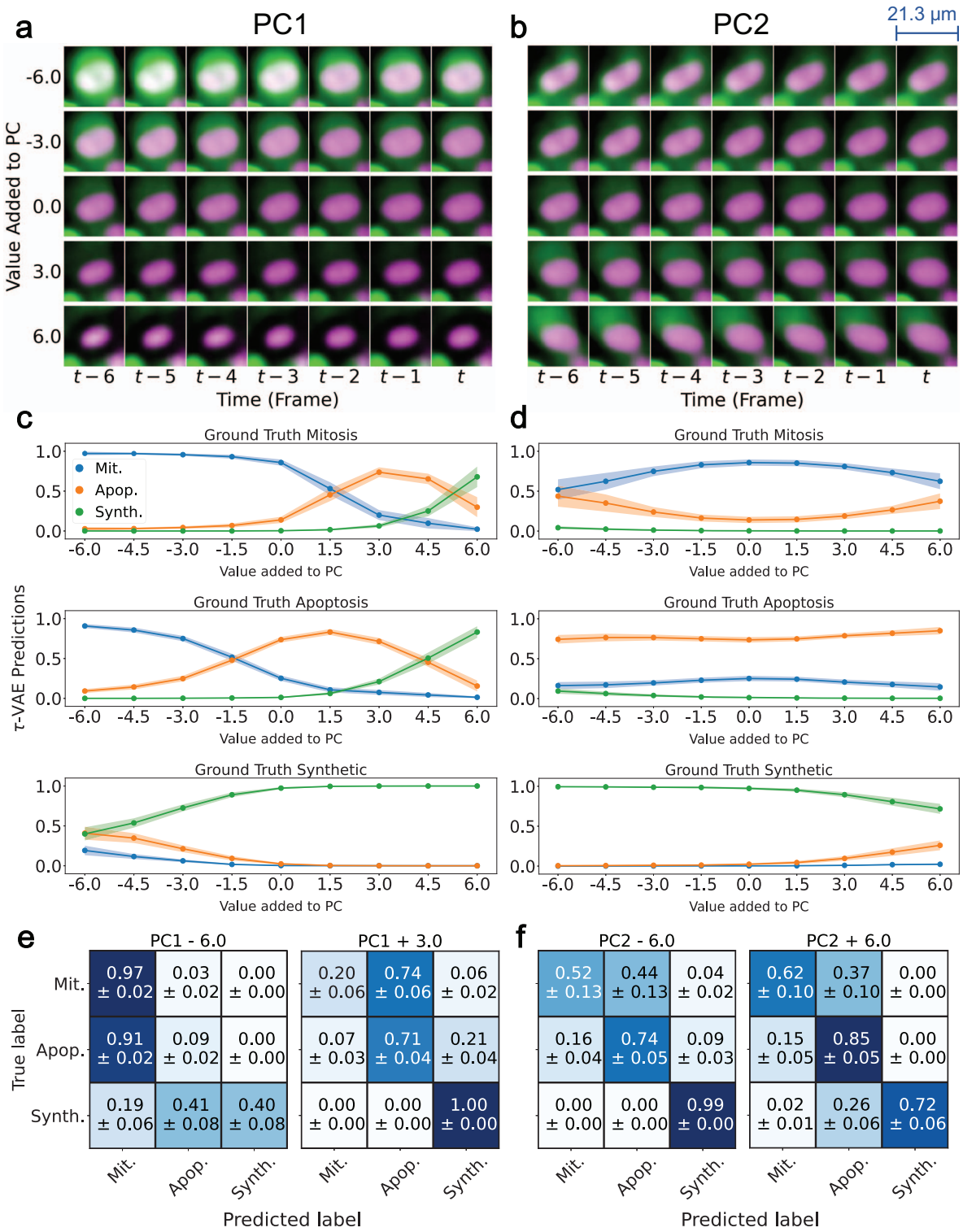

Figure 5. **PC shifting. a, b)** Decoded images, obtained by projecting the shifted PC representations back into latent space then passing the result through the decoder part of the β-VAE. Each row represents one constant value of a PC that has been added to each frame throughout the trajectory. **c, d)** Prediction outputs of the τ-VAE on the PC-shifted input data. Plotted are the proportions of test trajectories predicted as each class, with separate sub-figures associated with separate ground-truth labels. Translucent fill-in indicates the standard deviation across models. **e, f)** Confusion matrices at both extremes of the realistic range for each PC.

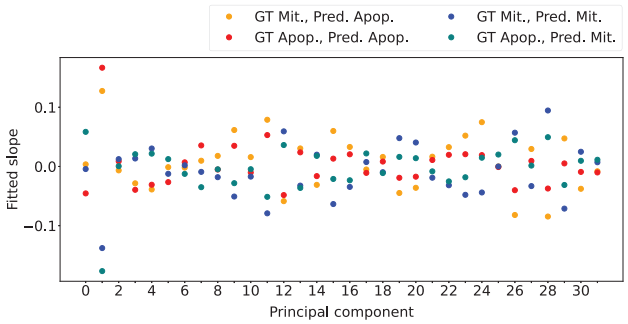

Figure 6. **Slopes of linear models fitted to the prediction curves by ordinary least-squares regression.** For each data-point, the mean value was considered. This calculation only used those data-points for which less than 20% of trajectories in the ground-truth class were predicted as synthetic, representing the "realistic range".

of throughout the entire trajectory, one could investigate whether apoptosis is dependent on the duration of crowding, as well as its severity. The possibilities are limited only by our ability to apply perturbations in a controlled manner. Furthermore, while the specific perturbations we use here have been tailored to cell imaging, the technique of virtual perturbation itself can be applied to any arbitrary context. For example, one can imagine audio-based perturbations (e.g., intensity or pitch shifting), text-based perturbations (e.g., word substitution) and so on.

That being said, it is important to consider whether the perturbations move the data out-of-distribution (OOD). DNNs trained to perform well within a specific distribution will not generally perform well outside of it. Hence, one should question the scientific value of perturbations that lead to OOD effects. At the same time, the detection of OOD perturbations (corresponding to synthetic predictions in our case) could be useful because it indicates those situations that are not observed in the training data.

Moving forward, this technique could serve as a useful addition to the arsenal of deep-learning based scientific methods, which already includes feature ablation [28], linear discriminant analysis [38], symbolic regression [32, 16] and more. One advantage of virtual perturbation is that it does not require any assumptions regarding the form of the downstream model[1], in the way that linear discriminant analysis prescribes a linear classifier, and symbolic regression searches for elegant mathematical expressions. In fact, one advantage of this technique is that it allows for investigation of a model's behavior without detailed knowledge of its internal mechanisms. This is achieved by recording statistically significant consequences of perturbation across a population with as great a diversity as possible.

Moreover, we sought not only to test the resilience of a particular causal relationship throughout the dataset, but also across several different models, each trained to high accuracy. In Section 1.2 we introduced the notion of an equal-performance "Rashomon" set G [26]. The existence of multiple models in this set would indicate that there are multiple internal mechanisms through which high performance could be obtained, only one of which would correspond to the natural phenomenon f. We therefore sought to discover those input-output relationships that were robust across different members of G. We did this by training and testing ten different models through 10-fold cross-validation, reporting small standard deviations in our results across models (Figs. 4 & 5). Of course, it would be impertinent to say that this robustness automatically proves that the relationship extends to f as well; however, it certainly suggests that this is the case. In other words, our study certainly suggests that in the case of mechanical cell competition, crowding is a strong determinant of cell fate, on the basis that ten models independently trained to predict cell fate have all leveraged PC1 as an important input feature. Future studies could explore this concept in a more rigorous manner, to discover those properties common to all members of G.

Researchers in various domains may use virtual perturbation to generate hypotheses for physical experimentation. By rapidly exploring the space of possible experiments, it would allow researchers to carry out only those that have a good chance of being useful. For example, had there been no prior experiments on the effect of crowding on scrib^{kd} cell apoptosis, the present study would have flagged this as a hypothesis that could be tested. Overall, we hope that this technique will serve others well across a variety of domains, and that it will contribute meaningfully to the exciting field of deep-learning based scientific investigation.

5. Acknowledgements

CJS was supported by a BBSRC LIDo studentship. ARL and GC wish to acknowledge support from BBSRC grant BB/S009329/1. We thank Kristina Ulicna and Marjan Famili for feedback on the manuscript.

References

[1] Assaf Arbelle and Tammy Riklin Raviv. Microscopy Cell Segmentation via Convolutional LSTM Networks, Jan. 2019. arXiv:1805.11247 [cs].

[2] Anna Bove, Daniel Gradeci, Yasuyuki Fujita, Shiladitya Banerjee, Guillaume Charras, and Alan R. Lowe. Local cellular neighborhood controls proliferation in cell competition. *Molecular Biology of the Cell*, 28(23):3215–3228, Nov. 2017.

[1] "Downstream" here refers to the model (or part thereof) that operates on the input *after* perturbation. Hence, while the effectiveness of PC1 manipulation does not depend on the downstream model (the TCN) it certainly does depend on the disentangled representation of the upstream β-VAE.

[3] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE, Apr. 2018. arXiv:1804.03599 [cs, stat].

[4] Nizam Ud Din and Ji Yu. Training a deep learning model for single-cell segmentation without manual annotation. *Scientific Reports*, 11(1):23995, Dec. 2021. Number: 1 Publisher: Nature Publishing Group.

[5] Meghan K. Driscoll and Assaf Zaritsky. Data science in cell imaging. *Journal of Cell Science*, 134(7):jcs254292, Apr. 2021.

[6] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, Alexander Dovzhenko, Olaf Tietz, Cristina Dal Bosco, Sean Walsh, Deniz Saltukoglu, Tuan Leng Tay, Marco Prinz, Klaus Palme, Matias Simons, Ilka Diester, Thomas Brox, and Olaf Ronneberger. U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16(1):67–70, Jan. 2019. Number: 1 Publisher: Nature Publishing Group.

[7] Donald Gillies. *Artificial Intelligence and Scientific Method*. Oxford University Press, Oxford, New York, Sept. 1996.

[8] Peter Goldsborough, Nick Pawlowski, Juan C. Caicedo, Shantanu Singh, and Anne E. Carpenter. CytoGAN: Generative Modeling of Cell Images, Dec. 2017. Pages: 227645 Section: New Results.

[9] Daniel Gradeci, Anna Bove, Giulia Vallardi, Alan R Lowe, Shiladitya Banerjee, and Guillaume Charras. Cell-scale biophysical determinants of cell competition in epithelia. *eLife*, 10:e61011, May 2021. Publisher: eLife Sciences Publications, Ltd.

[10] Shuyao Gu, Rachel M. Lee, Zackery Benson, Chenyi Ling, Michele I. Vitolo, Stuart S. Martin, Joe Chalfoun, and Wolfgang Losert. Label-free cell tracking enables collective motion phenotyping in epithelial monolayers. *iScience*, 25(7):104678, July 2022.

[11] Yuchen R. He, Shenghua He, Mikhail E. Kandel, Young Jae Lee, Chenfei Hu, Nahil Sobh, Mark A. Anastasio, and Gabriel Popescu. Cell Cycle Stage Classification Using Phase Imaging with Computational Specificity. *ACS Photonics*, 9(4):1264–1273, Apr. 2022. Publisher: American Chemical Society.

[12] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. Nov. 2016.

[13] Chetak Kandaswamy, Luís M. Silva, Luís A. Alexandre, and Jorge M. Santos. High-Content Analysis of Breast Cancer Using Single-Cell Deep Transfer Learning. *Journal of Biomolecular Screening*, 21(3):252–259, Mar. 2016.

[14] Mirae Kim, Soonwoo Hong, Thomas E. Yankeelov, Hsin-Chih Yeh, and Yen-Liang Liu. Deep learning-based classification of breast cancer cells using transmembrane receptor dynamics. *Bioinformatics (Oxford, England)*, 38(1):243–249, Dec. 2021.

[15] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, Dec. 2022. arXiv:1312.6114 [cs, stat].

[16] Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia. Rediscovering orbital mechanics with machine learning, Feb. 2022. arXiv:2202.02306 [astro-ph].

[17] Alan R. Lowe. Bayesian Tracker (btrack), May 2023. original-date: 2017-08-23T12:52:07Z.

[18] Golnaz Moallem, Adity A. Pore, Anirudh Gangadhar, Hamed Sari-Sarraf, and Siva A. Vanapalli. Detection of live breast cancer cells in bright-field microscopy images containing white blood cells by image analysis and deep learning. *Journal of Biomedical Optics*, 27(7):076003, July 2022.

[19] Mark Norman, Katarzyna A. Wisniewska, Kate Lawrenson, Pablo Garcia-Miranda, Masazumi Tada, Mihoko Kajita, Hiroki Mano, Susumu Ishikawa, Masaya Ikegawa, Takashi Shimada, and Yasuyuki Fujita. Loss of Scribble causes cell competition in mammalian cells. *Journal of Cell Science*, 125(Pt 1):59–66, Jan. 2012.

[20] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, Sept. 2016. arXiv:1609.03499 [cs].

[21] Anton Osokin, Anatole Chessel, Rafael E. Carazo Salas, and Federico Vaggi. GANs for Biological Image Synthesis, Sept. 2017. arXiv:1708.04692 [cs, stat].

[22] Wei Ouyang and Christophe Zimmer. The imaging tsunami: Computational opportunities and challenges. *Current Opinion in Systems Biology*, 4:105–113, Aug. 2017.

[23] Nick Pawlowski, Juan C. Caicedo, Shantanu Singh, Anne E. Carpenter, and Amos Storkey. Automating Morphological Profiling with Generic Deep Convolutional Networks, Nov. 2016. Pages: 085118 Section: New Results.

[24] Luca Rappez, Alexander Rakhlin, Angelos Rigopoulos, Sergey Nikolenko, and Theodore Alexandrov. DeepCycle reconstructs a cyclic cell cycle trajectory from unsegmented cell images using convolutional neural networks. *Molecular Systems Biology*, 16(10):e9474, Oct. 2020.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs].

[26] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the Existence of Simpler Machine Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, June 2022. arXiv:1908.01755 [cs, stat].

[27] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, Sept. 2015. arXiv:1506.04214 [cs].

[28] Christopher J. Soelistyo, Giulia Vallardi, Guillaume Charras, and Alan R. Lowe. Learning biophysical determinants of cell fate with deep neural networks. *Nature Machine Intelligence*, 4(7):636–644, July 2022. Number: 7 Publisher: Nature Publishing Group.

[29] Christoph Sommer, Rudolf Hoefler, Matthias Samwer, and Daniel W. Gerlich. A deep learning and novelty detection framework for rapid phenotyping in high-content screening. *Molecular Biology of the Cell*, 28(23):3428–3436, Nov. 2017.

[30] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[31] Luke Ternes, Mark Dane, Sean Gross, Marilyne Labrie, Gordon Mills, Joe Gray, Laura Heiser, and Young Hwan Chang. A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis. *Communications Biology*, 5(1):1–10, Mar. 2022. Number: 1 Publisher: Nature Publishing Group.

[32] Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, Apr. 2020. Publisher: American Association for the Advancement of Science.

[33] Kristina Ulicna, Laure T. L. Ho, Christopher J. Soelistyo, Nathan J. Day, and Alan R. Lowe. Convolutional Neural Networks for Classifying Chromatin Morphology in Live-Cell Imaging. *Methods in Molecular Biology (Clifton, N.J.)*, 2476:17–30, 2022.

[34] Kristina Ulicna, Giulia Vallardi, Guillaume Charras, and Alan R. Lowe. Automated Deep Lineage Tree Analysis Using a Bayesian Single Cell Tracking Approach. *Frontiers in Computer Science*, 3, 2021.

[35] Laura Wagstaff, Maja Goschorska, Kasia Kozyrska, Guillaume Duclos, Iwo Kucinski, Anatole Chessel, Lea Hampton-O'Neil, Charles R. Bradshaw, George E. Allen, Emma L. Rawlins, Pascal Silberzan, Rafael E. Carazo Salas, and Eugenia Piddini. Mechanical cell competition kills cells via induction of lethal p53 levels. *Nature Communications*, 7:11373, Apr. 2016.

[36] Roy Wollman and Nico Stuurman. High throughput microscopy: from raw images to discoveries. *Journal of Cell Science*, 120(21):3715–3722, Nov. 2007.

[37] Wenjin Yu, Yangyang Liu, Yunsong Zhao, Haofan Huang, Jiahao Liu, Xiaofeng Yao, Jingwen Li, Zhen Xie, Luyue Jiang, Heping Wu, Xinhao Cao, Jiaming Zhou, Yuting Guo, Gaoyang Li, Matthew Xinhu Ren, Yi Quan, Tingmin Mu, Guillermo Ayuso Izquierdo, Guoxun Zhang, Runze Zhao, Di Zhao, Jiangyun Yan, Haijun Zhang, Junchao Lv, Qian Yao, Yan Duan, Huimin Zhou, Tingting Liu, Ying He, Ting Bian, Wen Dai, Jiahui Huai, Xiyuan Wang, Qian He, Yi Gao, Wei Ren, Gang Niu, and Gang Zhao. Deep Learning-Based Classification of Cancer Cell in Leptomeningeal Metastasis on Cytomorphologic Features of Cerebrospinal Fluid. *Frontiers in Oncology*, 12, 2022.

[38] Assaf Zaritsky, Andrew R. Jamieson, Erik S. Welf, Andres Nevarez, Justin Cillay, Ugur Eskiocak, Brandi L. Cantarel, and Gaudenz Danuser. Interpretable deep learning uncovers cellular properties in label-free live cell images that are predictive of highly metastatic melanoma. *Cell Systems*, 12(7):733–747.e6, July 2021.

[39] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT. *Multimedia Tools and Applications*, 81(17):24265–24300, July 2022.