Discovering interpretable models of scientific image data with deep learning

Christopher J. Soelistyo   Alan R. Lowe
The Alan Turing Institute
{csoelistyo,alowe}@turing.ac.uk

Abstract

In this study, we demonstrate the possibility of finding interpretable, domain-appropriate models of biological images, and propose that such a strategy can be used to derive scientific insight in domains involving raw data. This is achieved by the novel, concerted application of existing methods, namely, disentangled representation learning, sparse deep neural network training and symbolic regression. We demonstrate their relevance to the field of bioimaging using a well-studied test problem of classifying cell states in microscopy data. We find that such methods can produce highly parsimonious models that achieve $\sim$ 98% *of the accuracy of black-box benchmark models, with a tiny fraction of the complexity, and greater domain-appropriateness, as tested by adversarial attacks. As such, we provide proof of concept that interpretable, high-performing models can be used to produce scientific explanations of some underlying biological phenomenon.*

1. Introduction

Advances in artificial intelligence have enabled data-driven modes of scientific discovery, where observations play a key role in an inductive process of theory construction. Deep neural networks (DNN) in particular present some unique strengths, owing to their high expressivity, which enables high performance on a broad range of tasks. Moreover, they can be trained using a general-purpose heuristic - gradient descent via backpropagation - and they are capable of effectively dealing with raw forms of data such as images and audio. This latter property makes them ideal for fields heavily dependent on the analysis of raw data, such as bio-imaging [28] and astrophysics [36].

However, DNNs also possess some properties that may be detrimental to scientific discovery. The most glaring is their complexity, which renders them inherently uninterpretable. Moreover, the high complexity of a DNN trained on a task would generally exceed the minimal function complexity required to conduct the task, in violation of Occam's razor. This also exposes them to the risk of "shortcut learning" [13], where the model learns to perform a task in undesirable ways - this may include the use of domain-inappropriate input features.

Hence, DNNs typically achieve high performance at the expense of interpretability and domain-appropriateness. The picture is one of an "accuracy-simplicity trade-off", where the former is gained at the expense of the latter. However, an alternate view holds that for a given task, there may exist a broad set of maximal performance models, a "Rashomon set" [3, 35], that includes both interpretable, domain-appropriate models as well as un-interpretable, domain-inappropriate ones. We aim to demonstrate the correctness of this latter view by finding high-performance, interpretable models for a well-studied test problem in bioimaging.

Several factors make bioimaging an ideal domain of application for deep-learning-based scientific discovery. The introduction of tools such as high-throughput microscopy has led to an explosion in the availability of bioimaging data [28, 42]. Furthermore, these data take the form of images, which typically include complex biological structures and a great deal of noise. They are therefore typically difficult to analyze without the aid of computational techniques [10]. Moreover, the sheer complexity of biological systems such as living cells often necessitates the use of tools that can extract high-level patterns and regularities, such as DNNs.

2. Prior work

Machine learning has been used extensively for the data-driven discovery of scientific rules. Early work focused on discovery of symbolic physical laws in tabular data, where input variables are associated with semantic labels. This includes the BACON and DALTON programs [21–23], and the series of genetic evolution algorithms developed afterward [9, 16, 20, 33, 40].

Deep learning and advances in computer vision have facilitated the use of raw data, such as images or audio, where the input variables (*e.g.*, pixels) do not carry semantic labels in themselves. Work here has largely involved representation learning, including the discovery of minimal state variable representations [6] or disentangled representations that can be used for downstream tasks [37, 38, 44]. Deep learning has been employed extensively in the cell biology field, which typically involves complex raw data [18, 26, 27].

Many approaches enforce domain-appropriateness by adding explicit constraints embodying prior knowledge. In physics, this has involving the enforcement of known physical laws [5, 17, 29] or the use of graph neural networks [24, 43] or neural production rules [15] to explicitly model interactions between physical entities.

Despite this progress, a gap remains in the development of interpretable, mathematically flexible, and minimal models of raw data. Existing applications of symbolic regression have relied on tabular data. Where raw data has been used, the models are either un-interpretable [6], restricted in mathematical form [44] or sensitive to domain-irrelevant aspects of the input [37]. Moreover, some widely used explanation methods such as SHAP [25], LIME [30] or GRAD-CAM [34] operate by creating local surrogate models of the underlying black box model; the target model itself remains un-interpretable [32]. While these explanations reveal some insight into the behavior of a black box model, they fall short of providing a reliable, comprehensive account of its operations [31].

3. Goal and strategy

The aim of the this study is to provide proof of concept that it is possible to obtain models of raw data that are simultaneously interpretable, domain-appropriate and high-performing. Moreover, we suggest that existing deep-learning methods are sufficient for the task, when applied in a novel framework. We assess our strategy on a test problem, taken from the field of bioimaging.

3.1. The test problem: classifying chromatin morphology in live-cell microscopy data

The scientific question on which we tested our models is, “what distinguishes a cell in interphase from one in metaphase?”, where these are distinct stages of the cell cycle. The input data for our models are single-cell images taken via fluorescence microscopy, where fluorescently tagged histone markers are used to enable visualization of chromatin. The target output is the associated cell state label: interphase or metaphase.

Prior knowledge informs us that the distinguishing characteristic relates to the organization of chromatin within the cells. When a cell is in interphase, the chromatin is distributed very diffusely around the nucleus. However, when it is in metaphase, the chromatin is aligned very sharply along an axis, in preparation for cell division. These differences are apparent in the microscopy data (see Fig. 7 for examples).

The image dataset comprising cells in each of these two stages, was acquired using cell culture, high-throughput fluorescence microscopy and automated cell tracking [2, 37, 39, 41]. In these experimental datasets, we cultured MDCK (Madin-Darby Canine Kidney) cells. Pixel intensities in the image dataset correspond to the density of chromatin.

Each image contains a central cell, as well as the neighboring cells around them, and a great deal of noise. We identify three domain constraints against which to evaluate our models.

1. The models use only information relevant to chromatin organization of the target cell (*i.e.*, not the neighboring cells). This arises from the definition of “cell state” as a property specific to the state of one cell.
2. Model outputs are invariant to transformations that affect the spatial orientation of our images (*e.g.*, rotations). This is because these factors in turn depend only on the spatial orientation of the microscope, and not on the underlying biological system itself.
3. Model outputs are insensitive to noise in the image.

Our dataset consisted of 3929 metaphase images and 4092 interphase images. We used a $90\% : 10\%$ split between our training and testing sets.

3.2. The strategy

We chose three main methods for increasing the parsimony and interpretability of our deep-learning-based models:

1. **Disentangled representation learning:** The discovery of a semantic latent representation, whose elements correspond to separate concepts. Representation learning models can transform raw data into semantically meaningful data, which can be used for downstream tasks such as classification. For this, we use a β-TCVAE [7].
2. **Sparse neural network training:** Training of minimally connected neural networks that select inputs discriminately and minimize the complexity of the learnt function.
3. **Symbolic regression:** Discovery of high-performing symbolic expression models, using the latent features deemed relevant by the sparse training procedure.

Our general approach is to train multiple models on the test problem, including some that use these methods and some that do not. We then analyze these models to assess the impact of these methods on the criteria of performance, interpretability and domain-appropriateness. For the latter criterion, we employ adversarial attacks to discover those image perturbations that can induce changes in classification, then compare those with our pre-established domain constraints (Sec. 3.1).

To this end, we train four different model types, of differing interpretability, on our test problem. In order of increasing interpretability, they are:

1. **Scheme 1: Dense CNN + Dense Head**: Dense convolutional neural network, followed by a dense classification head comprised of fully-connected layers.

2. **Scheme 2: β-TCVAE + Dense Head**: Convolutional β-TCVAE followed by a fully-connected dense classification head.
3. **Scheme 3: β-TCVAE + Sparse Head**: Convolutional β-TCVAE followed by a sparse classification head.
4. **Scheme 4: β-TCVAE + Symbolic expression**: Convolutional β-TCVAE, whose latent variables are related to the model output by a symbolic expression.

Scheme 1 models are completely un-interpretable; they act as a baseline for the interpretability assessment of Scheme 2-4 models. These latter models rely on a semantic latent space. Scheme 2 models use a highly complex classification function while Scheme 3 & 4 models use simpler functions. Crucially, Scheme 4 models are completely interpretable - they express the classification function as a mathematical expression based on the β-TCVAE latent representation. Scheme 3 models are interpretable as well, albeit, as we shall see, with more difficulty (see Sec. 5.3).

4. Methods

4.1. Total Correlation VAE

The Total Correlation Variational Autoencoder (β-TCVAE) [7] is a variant of the variational autoencoder (VAE) [19]. The VAE is a latent variable model that consists of an encoder network that compresses the input data into a probabilistic latent representation, and a decoder network that reconstructs the original input from this latent vector. The β-TCVAE is a variant of the VAE that has been designed specifically to produce disentangled latent spaces, where separate latent variables encode separate concepts. When applied on images, these may correspond to visual concepts such as size and shape.

We trained the model on roughly 2.1 million images randomly sampled from our microscope footage. We extracted 64×64 pixel crops around each cell, which corresponds to roughly 21.3μm along each side.

4.2. Sparsity: RigL

For sparse neural network training, we use a dynamic pruning algorithm known as *RigL* [11]. This algorithm is premised on the “lottery ticket hypothesis”, which states that dense neural networks will contain sub-networks that, when trained in isolation, can achieve test performance that matches the original dense network [12]. The general aim of sparse training algorithms is to identify this sub-network. *RigL* achieves this by dynamically pruning and re-growing connections at fixed intervals during training. When “pruned”, a connection weight is set to zero, and it ceases to update during training.

We adapt this algorithm by introducing a “warm-up” period at the beginning of training where the network is trained densely, and we implement two post-training pruning steps that remove unnecessary connections from the final network. Details can be found in the supplementary material.

4.3. Symbolic regression

Symbolic regression is a method to identify analytic expressions that approximate the output of an arbitrary function or dataset. We used PySR [8], an open-source symbolic regression package that runs a genetic evolutionary algorithm [20] to optimize symbolic expressions with respect to some fitness metric, with alternating rounds of mutation and tournament selection. This metric typically accounts for both the performance and complexity of the expressions.

4.4. Adversarial attacks

To assess the robustness of our classification networks, we implement a simple and efficient adversarial attack known as the Fast Gradient Sign Method (FGSM) [14]. This attack transforms some input data $\mathbf{x}$ with predicted class label y, such that the perturbed data $\tilde{\mathbf{x}}$ is classified by the model into another class $\tilde{y}$. FGSM forms $\tilde{\mathbf{x}}$ by adding to $\mathbf{x}$ some perturbation $\boldsymbol{\eta}$; *i.e.*, $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\eta}$. This perturbation is calculated based on the sign of the gradient of the loss function L with respect to the input $\mathbf{x}$:

$$\boldsymbol{\eta} = \epsilon \, \mathrm{sign}(\nabla_{\mathbf{x}} L(\boldsymbol{\theta}, \mathbf{x}, y)), \tag{1}$$

where ϵ is the pre-specified perturbation magnitude and $\boldsymbol{\theta}$ represents the network parameters.

5. Results

5.1. Disentangling image factors

When trained, the reconstructions produced by the β-TCVAE sufficiently captured the distinct morphological features of the cells in the dataset. Moreover, the β-TCVAE managed to extract disentangled latent features that were interpretable, including four that encode central cell morphology (Fig. 1). These include two that encode the size of the central cell chromatin signature (z_3 & z_{29}), and two that encode the eccentricity of this signature (z_{17} & z_{21}). Prior knowledge would inform us that only these four central cell morphology features would be relevant to cell state classification. The challenge was to assess whether our classification models conform to this assumption.

5.2. Cell state classification

For classification, we investigated the four model schemes introduced in Sec. 3.2. In Scheme 1, a Dense CNN reduces the input image to a feature vector, which is then processed by the dense fully-connected head. In Schemes 2, 3 & 4, the input is the latent representation of the image produced by the β-TCVAE. All models f reduce the input x to a single

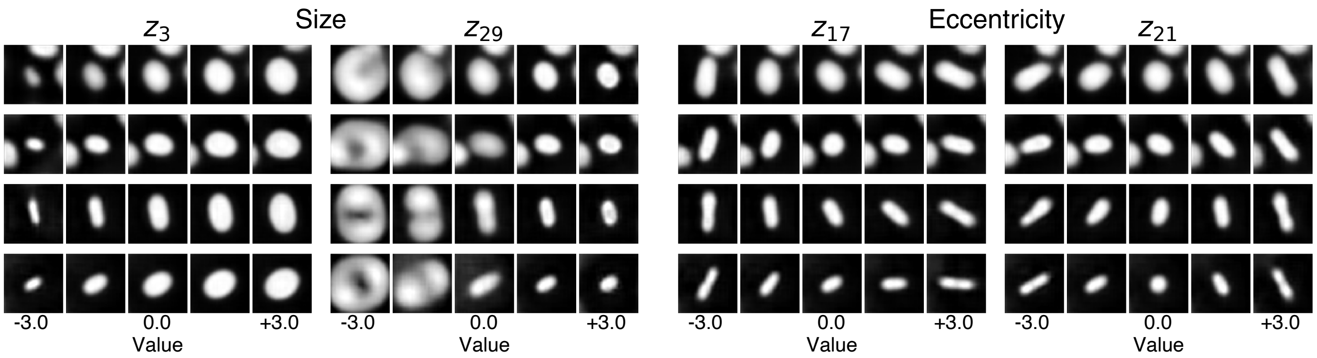


Figure 1. Latent-space traversals of variables that encode central cell chromatin organization.

output scalar $f(x)$, which is used to make the classification y following:

$$y = \begin{cases} \texttt{interphase}, & \text{if } f(x) < 0 \\ \texttt{metaphase}, & \text{if } f(x) \geq 0. \end{cases} \tag{2}$$

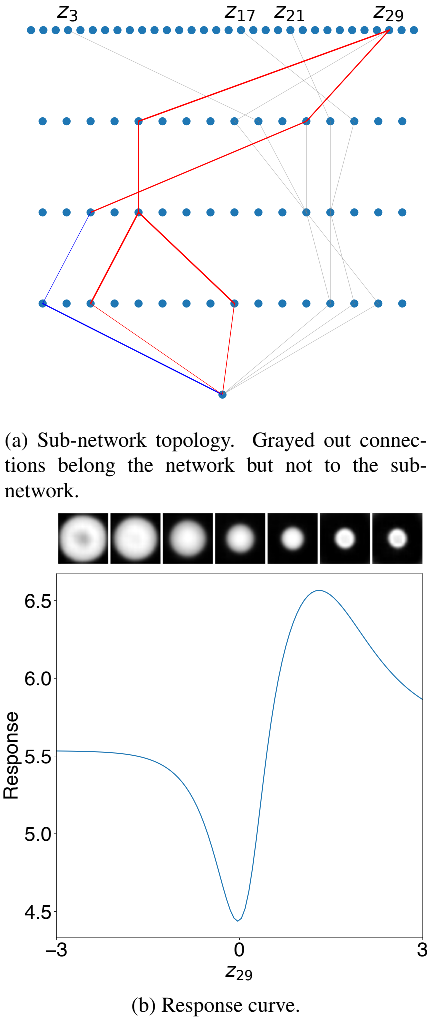
For training Scheme 3 models, we found the optimal sparsity level using the hyper-parameter search program OPTUNA [1]. For each scheme, we trained ten models (Tab. 1). More details on the acquisition of Scheme 4 models (involving symbolic expressions) can be found in Sec. 5.4. Strikingly, the enormous decrease in complexity of the classification head from Scheme 2 to 4 is accompanied only by a relatively minor decrease in performance, as measured by testing accuracy. For example, on average, Scheme 3 models attain 98% of the accuracy of Scheme 2 models, with only 2.2% of the active weight count and 2.1% of the expression size. Meanwhile, Scheme 4 models also attain about 98% of the accuracy of Scheme 2 models, but with only 0.2% of the expression size.

5.3. Sparse network analysis (scheme 3)

To assess the interpretability of our sparse networks, we chose our highest-performing Scheme 3 model (with 97.3% test accuracy) and inspected its behavior across its input space. The full topology of this model is shown in the supplementary material.

Strikingly, we find that the model has learnt that the minimal set of required input features corresponds exactly to the latent variables that encode central cell morphology (Fig. 1) and ignored those describing the neighborhood. In fact, this was true of all ten of our Scheme 3 models.

Our strategy for analyzing this network was to decompose it into a few sub-networks then study the behavior of their outputs (their "response") across their input spaces. This is only possible due to the sparsity of the overall network, which enables decomposition into a sensible number of sub-networks with low-dimensional input spaces. Sub-network 1 is shown in Fig. 2, while sub-networks 2, 3 & 4 are shown in Figs. 3 and 4.



(a) Sub-network topology. Grayed out connections belong the network but not to the sub-network.

(b) Response curve.

Figure 2. Sub-network 1. This sub-network responds to nuclear size.

Scheme	Encoder	Head	No. of head weights	Head expression size	Accuracy
1	CNN	Dense	1040	9697	$99.7 \pm 0.1\%$
2	VAE	Dense	1040	8641	$99.0 \pm 0.1\%$
3	VAE	Sparse	23 ± 2	180 ± 20	$97.0 \pm 0.2\%$
4	VAE	Symbolic	N/A	17 ± 7	$97.4 \pm 0.2\%$

Table 1. Testing performance across ten models within each scheme. Errors represent the standard deviation across ten models from each scheme. No. of head weights is the number of active connection weights. Head expression size is the number of nodes in the expression tree equivalent to the model concerned. For Scheme 1-3, head expression size is calculated according to a method outlined in the supplementary material.

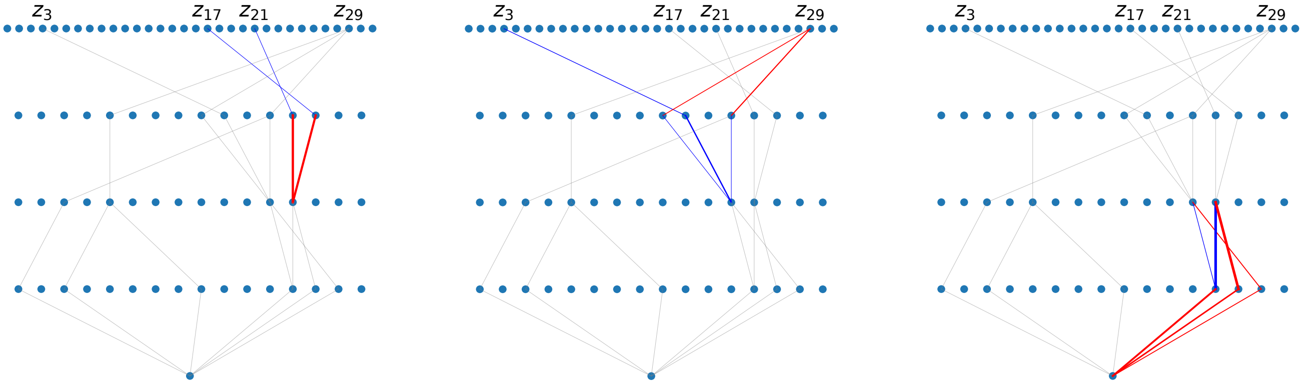


Figure 3. Topologies for sub-networks 2, 3, & 4 respectively (left to right). Grayed out connections belong the network but not to the sub-network.

Sub-network 1 accepts only the cell size variable z_{29}. The response curve reveals non-monotonic overall behavior; however, in the vicinity of $z_{29} = 0$, higher z_{29} monotonically elicits a higher response. Therefore, this sub-network responds to cell size in a straightforward fashion. However, we observe that its response is wholly constrained to the positive region, suggesting that this sub-network on its own cannot be decisive for classification.

Sub-network 2 accepts both size variables (z_3 & z_{29}) as input, and its response can be interpreted as a direct measure of cell size; it generally increases with increasing z_3 and decreasing z_{29}.

Sub-network 3 accepts both eccentricity variables as input, and its response can be interpreted as an orientation-independent measure of eccentricity; it is greatest as $z_{17}, z_{21} = 0$, which corresponds to maximal roundness, and decreases in radial fashion from the origin. Hence, the response behaves similarly with respect to both of the orientation-sensitive inputs z_{17} & z_{21}.

Sub-network 4 receives the outputs of both sub-networks 2 & 3 (hereafter, "z_{size}" and "z_{round}" respectively) as inputs, and its response serves as a contribution to the response of the overall network. We observe that the response is most sensitive to z_{size}; for any value of z_{round}, changes in z_{size} are sufficient to determine the classification. Sensitivity to z_{round} is comparably smaller. Finally, we observe that the response of sub-network 4 is pre-dominantly negative; hence in most cases, the output of sub-network 1 is required to push the final output into the positive range, which entails classification of metaphase.

From our brief study of this particular sparse network, we can therefore list some of its learned insights about the underlying system:

1. **Cell state is determined primarily by cell size**: Eccentricity terms z_{17} & z_{21} influence the final output only through z_{round}, whose impact is marginal relative to that of z_{size}. Moreover, in most cases, the output of sub-network 1 (which is dependent only on z_{29}) is required to push the final output into the positive region.
2. **Metaphase cells tend to be smaller than interphase cells**: The network response is typically increased by increasing z_{29} and decreasing z_3.
3. **Metaphase cells tend to be more eccentric than interphase cells**: For any fixed value of z_{size}, decreasing z_{round} will increase the response value of sub-network 4 in most cases.
4. **Cell state is independent of spatial orientation**: z_{17} & z_{21} are treated virtually equally. These variables influ-

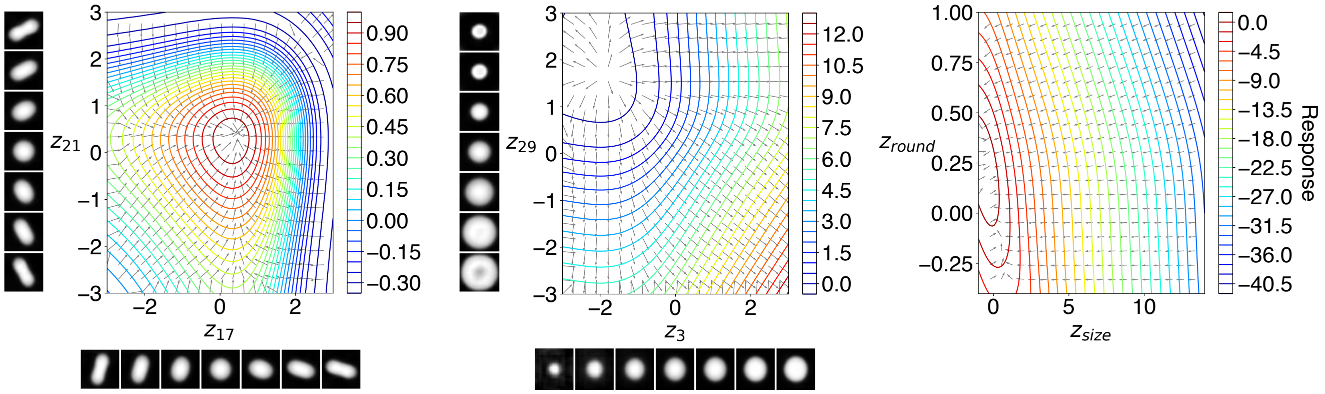


Figure 4. Response maps for sub-networks 2, 3, & 4 respectively (left to right). Colored contour lines represent the response value at each point in the latent sub-space. Gray arrows represent the gradient of the response value. Images are decoded traversals in latent space along the axis of each latent variable. In the sub-network 4 map (right), z_{round} and z_{size} refer to the output of sub-networks 2 & 3 respectively.

ence the final network output only through the response of sub-network 3 ("z_{round}"), and its response map is symmetric about the $z_{17} = z_{21}$ line.

In summary, we were able to gain significant insight into the behavior of our model due to its sparsity, which enabled us to decompose it into several sub-networks whose response functions can be studied in isolation. This in turn enabled us to derive some insights learnt by the model, which largely adhere to prior domain knowledge.

Nevertheless, analyzing sparse networks can be an onerous task, especially for domains or problems more complex than the one presented to our models here. Moreover, our analysis revealed some features of the model that we suspect reflects unnecessary complexity, such as the non-monotonic response dependence of sub-network 4 on z_{round} (Fig. 4) or that of sub-network 1 on z_{29} (Fig. 2b). Such behaviors may reflect the data, but may also reflect accidental biases introduced by the model architecture.

In Sec. 5.4, we explore how symbolic regression can address these problems by further reducing the complexity of the model.

5.4. Decision boundary discovery (scheme 4)

Our final strategy to reduce the complexity of the model is to find simple analytic expressions that produce interpretable decision boundaries in input space, using symbolic regression. This step benefits heavily from the reduction of the input space from 32 to 4 dimensions achieved by our sparse networks. This is because genetic algorithms (GAs) - such as that used by PySR - tend to scale poorly with the number of input variables, owing to the fact that linearly increasing the number of input variables would exponentially increase the number of possible expression trees [8].

Here, we analyze the symbolic expression model that captured the best balance between accuracy and simplicity, named "Exp. H1". The expression associated with this model is $z_{29}(z_{17}^2 + z_{21}^2) - e^{e^{z_3}}$. The rest of the expressions obtained can be found in the supplementary material.

What is immediately apparent about Exp. H1 (Fig. 5; test accuracy of 97.6%) is that the combined eccentricity term $(z_{17}^2 + z_{21}^2)$ can never be negative, regardless of the values of z_{17} and z_{21}. Therefore, it appears to function as a weighting term, modulating the balance between the two size components z_{29} and $e^{e^{z_3}}$. The double-exponent $e^{e^{z_3}}$ rises so sharply for $z_3 > 1$ so as to impose a virtual veto on any classifications of metaphase (Fig. 6). Meanwhile, for $z_3 < -1$, this term changes little, and metaphase classifications are allowed for $z_{29} > 0$, depending on the value of $(z_{17}^2 + z_{21}^2)$. In the range $-1 < z_3 < 1$, the role of this eccentricity weighting term becomes more decisive.

Again, we can interpret the principles that the model has learnt about the underlying biological system, many of which are shared with our Scheme 3 model.

1. **Cell state is determined primarily by cell size**: For sufficiently high values of z_3 (indicating large cell size), metaphase classifications are virtually impossible, regardless of the values of the other variables. Moreover, since the expression $-e^{e^{z_3}}$ is always negative, a sufficient value of z_{29} is required for metaphase classification.
2. **Metaphase cells tend to be smaller than interphase cells**: The output is monotonically increased by increasing z_{29} and decreasing z_3.
3. **Metaphase cells are eccentric**: At $z_{17}, z_{21} = 0$, indicating perfect roundness, classifications of metaphase are impossible. Beyond that, in the $z_{29} > 0$ region, higher $(z_{17}^2 + z_{21}^2)$, *i.e.*, higher eccentricity, increases the output.
4. **Cell state is independent of spatial orientation**: z_{17} & z_{21} are treated perfectly equally; moreover, Exp. H1 considers only the squares, and therefore magnitudes, of these terms.

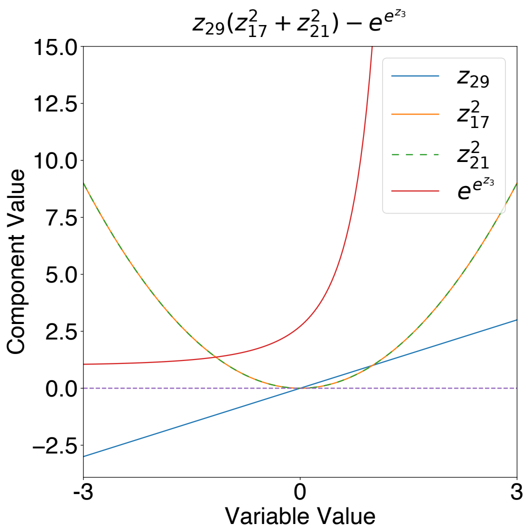


Figure 5. Exp. H1: component value mapped against the value of their relevant input variables. The purple dashed line represents a value of zero.

5.5. Counterfactual examples

To further test the domain-appropriateness of our models, we applied adversarial attacks on each model to find counterfactual examples. The particular attack we used, the Fast Gradient Sign Method (Sec. 4.4), perturbs the input by taking a jump within input space in the direction of the loss gradient, thereby worsening the performance of the model. The idea here is to find the minimal changes to the original input (Fig. 7) required to reverse the classification. By inspecting the nature of these modifications, we can study the sensitivity of our models to various aspects of the input. We analyze these results in light of the domain constraints outlined in Sec. 3.1.

We conducted two types of adversarial attack: within image space and within latent space. Image-based attacks were implemented on Scheme 1-4 models while latent-based attacks were implemented on Scheme 2-4 models only, given that Scheme 1 models do not possess an interpretable latent space.

Full results are shown in the supplementary material. Here, we highlight three main findings.

1. **Scheme 1 models are highly sensitive to noise.** Image-based perturbations are widely dispersed around the frame and do not lead to any differences in morphology significant enough to be captured by the β-TCVAE encoding (Fig. 8).
2. **Scheme 3 & 4 models are specifically sensitive to chromatin morphology.** Metaphase → Interphase image-based perturbations depress pixel values associated with the metaphase chromatin signature in order to simulate a more diffuse signature. Meanwhile, Interphase → Metaphase perturbations "carve out" a pill-shaped region at the center of the cell by depressing all other regions where chromatin is present. Similar effects are achieved by latent-based perturbations (Fig. 9).
3. **Scheme 2 models are sensitive to neighborhood density.** We obtained this result by carrying out latent-based perturbations on Scheme 2 models that specifically leave unchanged those latent variables that encode central cell morphology (z_3, z_{17}, z_{21} & z_{29}). We found that it was possible to reverse the classification by altering only those that encode aspects of the background. Interestingly, we found that Interphase → Metaphase perturbations decrease neighborhood density through the size reduction or outright removal of neighborhood cells, while Metaphase → Interphase perturbations do the reverse (Fig. 10). We suggest that this arises from the statistical dependence between neighborhood crowding and cell state. MDCK cells follow an "adder" model of size control, meaning that they divide after they have added a certain volume to their initial size [4]. Hence, cells in metaphase, which by definition are dividing, tend to be larger, and so would have displaced neighboring cells from their immediate vicinity. While this approach may increase model performance, it represents a clear "shortcut" strategy [13] as it contradicts domain knowledge. Hence, the model could assign opposite classifications to two images with the same central cell morphology, but different neighborhoods - a failure that clearly constitutes domain-inappropriate behavior.

6. Discussion

The primary conclusion of this study is that, for the test problem considered, it is possible to train interpretable, highly performant models that suffer only a minimal decrease in test accuracy in exchange for a profound increase in interpretability, with respect to our baseline Scheme 1 & 2 models. Indeed, on average, Scheme 3 classification models are able to capture 98% of the test performance of Scheme 2 models with only 2.2% of the active weight count and 2.1% of the expression size while Scheme 4 models could achieve 98% of the performance with only 0.2% of the expression size (Tab. 1).

Furthermore, we managed to form an interpretable, semantic representation of the raw image data, and train minimally complex classification models on the latent representation, using only techniques that exist in the current machine learning literature (β-TCVAE, *RigL* and symbolic regression). The novelty of this study is in their concerted application within a specified strategy. This suggests that in the current state of the field, there is much room for the application of interpretable deep learning within a broad range of scientific domains, even those that utilize raw data.

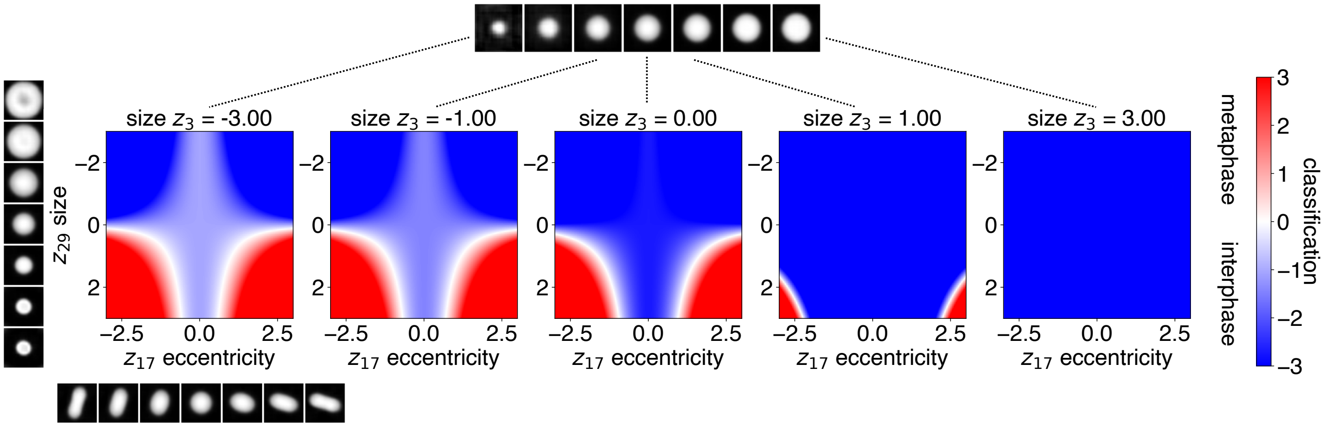


Figure 6. The decision boundary of Exp. H1, plotted for varying values of z_3, z_{17} & z_{29} within the typical range of values [-3, 3]. The value of z_{21} is held constant at 0 for the purposes of clarity.

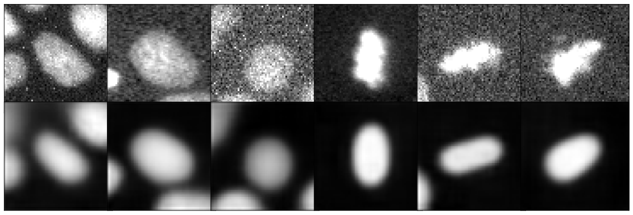


Figure 7. Un-perturbed image examples. **Top:** Image. **Bottom:** β-TCVAE reconstruction. **Left to Right:** Three interphase examples, then three metaphase examples.

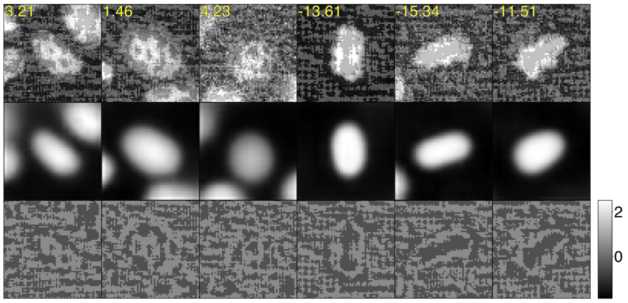


Figure 8. Image-based attacks on Scheme 1 models at $\epsilon = 0.5$. **Top:** Perturbed image. **Middle:** β-TCVAE reconstruction. **Bottom:** Perturbation. Numbers shown are the post-attack classification scores.

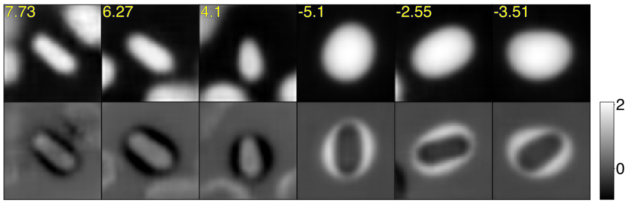


Figure 9. Latent-based attacks on Scheme 4 models at $\epsilon = 1.0$. **Top:** β-TCVAE decoding of the perturbed latent vector. **Bottom**: Perturbation in image space.

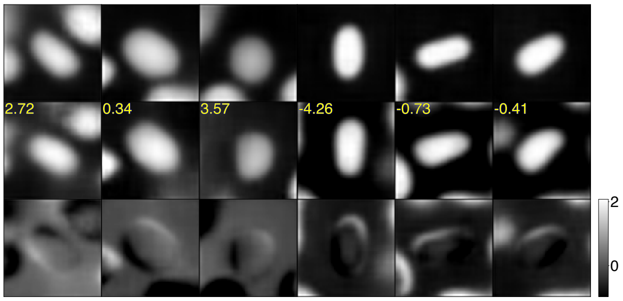


Figure 10. Latent-based attacks at $\epsilon = 1.0$, excluding z_3, z_{17}, z_{21} & z_{29}. **Top**: β-TCVAE decoding of the original latent vector. **Middle**: Decoding of the perturbed latent vector. **Bottom**: Perturbation in image space.

Moving forward, the challenge would be to apply these methods to systems of greater complexity. The test problem chosen here was relatively simple; the images contain few factors of variation, and the task was a straightforward binary classification between two classes easily distinguishable by the trained human eye. Could we extend such methods to more complex systems, and to tasks involving multi-class classification, or regression? We suggest that while such cases may require further design work, the problems are primarily technical and not fundamental.

In summary, this work has demonstrated the possibility of free-form scientific induction under constraint, using deep neural networks and interpretability techniques. There is much room for further exploration, and it is exciting to ponder the extent to which we can imbue machines with whatever it is that underlies our capacity for science.

Acknowledgements. This work was funded by The Alan Turing Institute under EPSRC grant EP/N510129/1, and supported by Baskerville under EPSRC Grant EP/T022221/1.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework, 2019. arXiv:1907.10902 [cs, stat]. 4

[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, 2017. 2

[3] Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215, 2001. Publisher: Institute of Mathematical Statistics. 1

[4] Clotilde Cadart, Sylvain Monnier, Jacopo Grilli, Pablo J. Sáez, Nishit Srivastava, Rafaele Attia, Emmanuel Terriac, Buzz Baum, Marco Cosentino-Lagomarsino, and Matthieu Piel. Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. *Nature Communications*, 9(1):3275, 2018. Number: 1 Publisher: Nature Publishing Group. 7

[5] Henry Chan, Youssef S. G. Nashed, Saugat Kandel, Stephan Hruszkewycz, Subramanian Sankaranarayanan, Ross J. Harder, and Mathew J. Cherukara. Real-time 3D Nanoscale Coherent Imaging via Physics-aware Deep Learning, 2020. arXiv:2006.09441 [cond-mat, physics:physics]. 2

[6] Boyuan Chen, Kuang Huang, Sunand Raghupathi, Ishaan Chandratreya, Qiang Du, and Hod Lipson. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2(7):433–442, 2022. Number: 7 Publisher: Nature Publishing Group. 1, 2

[7] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders, 2019. arXiv:1802.04942 [cs, stat]. 2, 3

[8] Miles Cranmer. PySR: High-Performance Symbolic Regression in Python and Julia, 2023. original-date: 2020-09-14T11:16:09Z. 3, 6

[9] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering Symbolic Models from Deep Learning with Inductive Biases, 2020. arXiv:2006.11287 [astro-ph, physics:physics, stat]. 1

[10] Meghan K. Driscoll and Assaf Zaritsky. Data science in cell imaging. *Journal of Cell Science*, 134(7):jcs254292, 2021. 1

[11] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the Lottery: Making All Tickets Winners, 2019. 3

[12] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, 2019. arXiv:1803.03635 [cs]. 3

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. Number: 11 Publisher: Nature Publishing Group. 1, 7

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, 2015. arXiv:1412.6572 [cs, stat]. 3

[15] Anirudh Goyal, Aniket Didolkar, Nan Rosemary Ke, Charles Blundell, Philippe Beaudoin, Nicolas Heess, Michael Mozer, and Yoshua Bengio. Neural Production Systems: Learning Rule-Governed Visual Dynamics, 2022. arXiv:2103.01937 [cs, stat]. 2

[16] Roger Guimerà, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel Miranda, Jordi Pallarès, and Marta Sales-Pardo. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 6(5):eaav6971, 2020. Publisher: American Association for the Advancement of Science. 1

[17] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. Number: 6 Publisher: Nature Publishing Group. 2

[18] Jacob C. Kimmel, Andrew S. Brack, and Wallace F. Marshall. Deep Convolutional and Recurrent Neural Networks for Cell Motility Discrimination and Prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(2):562–574, 2021. 2

[19] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, 2022. arXiv:1312.6114 [cs, stat]. 3

[20] John R. Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, 4(2):87–112, 1994. 1, 3

[21] Pat Langley. Data-Driven Discovery of Physical Laws. *Cognitive Science*, 5(1):31–54, 1981. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6708.1981.tb00869.x. 1

[22] Pat Langley, Herbert Simon, Gary Bradshaw, and Jan Zytkow. *Scientific Discovery: Computational Explorations of the Creative Process*. The MIT Press, Cambridge, Massachusetts, 1992.

[23] Patrick W. Langley. BACON: a production system that discovers empirical laws. In *Proceedings of the 5th international joint conference on Artificial intelligence - Volume 1*, page 344, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc. 1

[24] Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia. Rediscovering orbital mechanics with machine learning, 2022. arXiv:2202.02306 [astro-ph]. 2

[25] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, 2017. arXiv:1705.07874 [cs, stat]. 2

[26] Yukiko Nagao, Mika Sakamoto, Takumi Chinen, Yasushi Okada, and Daisuke Takao. Robust classification of cell cycle phase and biological feature extraction by image-based deep learning. *Molecular Biology of the Cell*, 31(13):1346–1354, 2020. Publisher: American Society for Cell Biology (mboc). 2

[27] Shori Nishimoto, Yuta Tokuoka, Takahiro G. Yamada, Noriko F. Hiroi, and Akira Funahashi. Predicting the future direction of cell movement with convolutional neural networks. *PLOS ONE*, 14(9):e0221245, 2019. Publisher: Public Library of Science. 2

[28] Wei Ouyang and Christophe Zimmer. The imaging tsunami: Computational opportunities and challenges. *Current Opinion in Systems Biology*, 4:105–113, 2017. 1

[29] G. P. Purja Pun, R. Batra, R. Ramprasad, and Y. Mishin. Physically informed artificial neural networks for atomistic modeling of materials. *Nature Communications*, 10(1):2339, 2019. Number: 1 Publisher: Nature Publishing Group. 2

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier, 2016. arXiv:1602.04938 [cs, stat]. 2

[31] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. Number: 5 Publisher: Nature Publishing Group. 2

[32] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16(none):1–85, 2022. Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada. 2

[33] Michael Schmidt and Hod Lipson. Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324(5923):81–85, 2009. Publisher: American Association for the Advancement of Science. 1

[34] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. arXiv:1610.02391 [cs]. 2

[35] 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, 2022. arXiv:1908.01755 [cs, stat]. 1

[36] Michael J. Smith and James E. Geach. Astronomia ex machina: a history, primer and outlook on neural networks in astronomy. *Royal Society Open Science*, 10(5):221454, 2023. Publisher: Royal Society. 1

[37] 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, 2022. Number: 7 Publisher: Nature Publishing Group. 2

[38] Christopher J. Soelistyo, Guillaume Charras, and Alan R. Lowe. Virtual perturbations to assess explainability of deep-learning based cell fate predictors. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 3973–3982. IEEE, 2023. 2

[39] Christopher J. Soelistyo, Kristina Ulicna, and Alan R. Lowe. Machine learning enhanced cell tracking. *Frontiers in Bioinformatics*, 3, 2023. 2

[40] Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. Publisher: American Association for the Advancement of Science. 1

[41] 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. 2

[42] Roy Wollman and Nico Stuurman. High throughput microscopy: from raw images to discoveries. *Journal of Cell Science*, 120(21):3715–3722, 2007. 1

[43] Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, 2018. arXiv:1710.10324 [cond-mat]. 2

[44] 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, 2021. 2