HAL
open science

What machine learning can do for developmental biology

Paul Villoutreix

To cite this version:

Paul Villoutreix. What machine learning can do for developmental biology. Development (Cambridge, England), 2021, 148 (1), pp.dev188474. 10.1242/dev.188474 . hal-03200144

HAL Id: hal-03200144

https://hal.science/hal-03200144v1

Submitted on 19 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

© 2021. Published by The Company of Biologists Ltd | Development (2021) 148, dev188474. doi:10.1242/dev.188474

SPOTLIGHT

What machine learning can do for developmental biology

Paul Villoutreix

ABSTRACT

Developmental biology has grown into a data intensive science with the development of high-throughput imaging and multi-omics approaches. Machine learning is a versatile set of techniques that can help make sense of these large datasets with minimal human intervention, through tasks such as image segmentation, super-resolution microscopy and cell clustering. In this Spotlight, I introduce the key concepts, advantages and limitations of machine learning, and discuss how these methods are being applied to problems in developmental biology. Specifically, I focus on how machine learning is improving microscopy and single-cell 'omics' techniques and data analysis. Finally, I provide an outlook for the futures of these fields and suggest ways to foster new interdisciplinary developments.

KEY WORDS: Artificial intelligence, Big data, Machine learning, Neural networks

Introduction

Developmental biology has undergone considerable transformations in the past 10 years. It has moved from the study of developmental genetics, which was mainly focused on the molecular scale, to an integrative science incorporating dynamics at multiple scales; from single molecules, to cells, to tissues. The development of new microscopy techniques, such as light-sheet microscopy (McDole et al., 2018) or high-resolution microscopy (Liu et al., 2018) has enabled the study of entire embryos and single molecular processes in the context of entire tissues. In parallel, high-throughput sequencing techniques have opened new avenues for the study of gene expression in developing embryos at a whole-genome scale (Briggs et al., 2018; Wagner et al., 2018; Farrell et al., 2018). These techniques generate terabytes of data and description in high-dimensional spaces; for this reason, it has become difficult to explore this wealth of data by hand. Therefore, we need to harness automated methods from computer science and statistics to explore these datasets, and extract the key features in a digestible manner. This requires asking the right questions and identifying the right approaches to answer them.

Concurrently, computer science is being revolutionized by machine learning, a subfield of artificial intelligence that implements inference algorithms (statistical inference methods made possible by significant computer power and large datasets) with minimal human intervention (Mohri et al., 2018) (Fig. 1). Machine learning can be broadly divided into supervised, unsupervised and reinforcement learning (Fig. 2), and aims at solving the following common tasks: classification, regression, ranking, clustering and dimensionality reduction (sometimes known as manifold learning) (Mohri et al., 2018) (Fig. 3). Following a rich history of methods applied to developmental problems, such as genetic algorithms to infer evolutionary-developmental (evo-devo) relationships (Azevedo et al., 2005), deep learning is a popular example of machine learning (Box 1; Fig. 1). Although deep learning was invented many years ago, it has come to full performance by taking advantage of large annotated datasets and the use of new hardware, such as the graphical processing unit (GPU). Since then, there has been a long line of very impressive successes of machine learning in a large variety of fields. One such milestone was the victory of the computer AlphaGo against the world champion of the strategic game Go (Silver et al., 2016), but beyond these technological achievements, how can machine learning really advance scientific fields?

It would be unrealistic to expect an exhaustive picture of the field of machine learning. However, in this Spotlight, I give a brief overview of this fast-growing, evolving field and explore the use of these different methods through two main axes: (1) how they can help improve microscopy techniques, and (2) how they can be used to analyze the large datasets produced by single-cell omics techniques.

Microscopy

Machine learning has been particularly instrumental in the field of fluorescent microscopy. Deep learning in particular has been designed for, and is extremely well adapted to, the study of images (Box 1). Here, I show how machine learning and deep learning helps to improve processing of microscopy data.

Image segmentation

Image segmentation is the process of automatically identifying regions of interest in an image and is an important part of quantification of microscopy images in developmental biology. Segmentation usually involves creating digital reconstructions of the shape of nuclei or cell membranes, which can then be used to quantify various features of these objects, such as their volume and shape index, as well as additional signals such as the expression of a given gene within these regions. For several years, the question of cell segmentation was treated using more or less explicit models, which means that they would include information *a priori* of the system under study. Typical examples include image intensity thresholding or active contours that incorporate a model of the continuity of the membranes being segmented. However, the question of image segmentation can be restated as a pixel classification task (Fig. 3A): does a pixel belong to an object or not? This leads to more generic models based on deep learning. Two recent approaches, StarDist and Cellpose, have been particularly successful at segmentation in 2D and 3D (Schmidt et al., 2018; Weigert et al., 2020; Stringer et al., 2020). These approaches are part of the family of supervised algorithms (Fig. 2A). Although very successful, they require large datasets of labeled data to train the models, which can be time consuming and costly to establish. However, a recent study suggests that it is possible to train segmentation models without needing labeled data, leading to an unsupervised learning algorithm (Hollandi et al., 2020) (Fig. 2B).

Large-scale screens and tracking

In addition to image segmentation, machine learning can be applied to *in toto* movies of developing embryos and high-throughput image

LIS (UMR 7020), IBDM (UMR 7288), Turing Center For Living Systems, Aix-Marseille University, 13009, Marseille, France.

*Author for correspondence (paul.villoutreix@univ-amu.fr)

P.V., 0000-0002-6333-5735

Development (2021) 148, dev188474. doi:10.1242/dev.188474

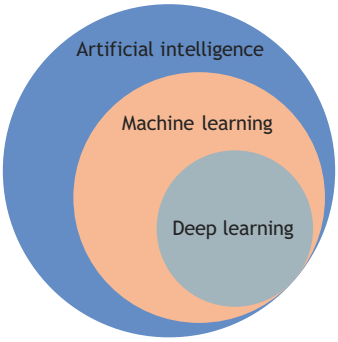

Fig. 1. Hierarchy of theories within the field of artificial intelligence. Artificial intelligence is the general field of computer science that aims to develop intelligent machines. Machine learning falls within the umbrella of artificial intelligence and is concerned with the task of making accurate predictions from very large datasets with minimal human intervention. Deep learning is a family of models that has been extremely successful at various tasks of machine learning, in particular with images.

data acquisition techniques for automatic phenotyping, cell tracking and cell lineage reconstruction. Large phenotypic screens can be performed automatically on large sets of images using supervised deep learning algorithms (Fig. 2A); these are called image classification or trajectory classification tasks (Khosravi et al., 2019; Zaritsky et al., 2020 preprint) (Fig. 3A). Cell tracking relates to the problem of object tracking, which is a widespread challenge in the image analysis field; several methods have been directly transferred to fluorescent microscopy images (Moen et al., 2019). Lineage reconstruction, however, is more specific to developmental biology because the tracked objects are dividing; dividing events can be detected with a supervised deep learning method (McDole et al., 2018). The precise history of these cell divisions can be obtained from time-lapse movies, cell barcoding (which is a molecular way of tagging cells of a similar descent, and consequently enabling the reconstruction of their division history) or a combination of these approaches. Once the cell lineage data are obtained, one can investigate the developmental origin of various tissues.

Image resolution

When considering the question of localizing individual molecules, one can turn to super-resolution microscopy, which has benefitted spectacularly from machine learning (Ouyang et al., 2018; Belthangady and Royer, 2019). To improve image resolution, one can train the model on a set of high-resolution images coupled with their less-well-resolved counterparts, as demonstrated by the method ANNA-PALM (Ouyang et al., 2018). Through this supervised approach, the model learns the mapping that converts a low-resolution image to a high-resolution image. Once learned, one can then acquire low-resolution images and improve them with the model to predict the high-resolution image. Here, the algorithm solves a regression task by predicting the value of individual pixels varying continuously (their intensity value), which is different from segmentation that instead aims to classify pixels (Fig. 3A,B). The method CARE has a similar conceptual approach to predict a de-noised image from a noisy one (Weigert et al., 2018). These types of approaches are supervised learning techniques and hence rely on high quality labeled datasets. It should be noted that, because they are generating new artificial images, they are prone to potential errors known as ‘hallucinations’, which consist of artefacts generated when improving images without basis in the original images.

Label-free imaging

Obtaining fluorescently labeled images is a long and costly process, particularly when using immuno-based staining. Some approaches have been proposed to predict labels from another less expensive type of microscopy such as transmitted-light microscopy using supervised deep learning techniques (Ounkomol et al., 2018; Christiansen et al., 2018). In these examples, the model is learned from pairs of images acquired in two modalities: the first modality being the transmitted-light microscopy, and the other, the labeled one. Once the model is learned, a new fluorescently labeled image can be predicted from an image containing only the first modality. Although very powerful, this label-free imaging framework works only on a limited number of labels (those that correlate with the first modality) and is thus not yet widely usable in developmental biology.

Data integration

When many images are produced with various modalities, how can one integrate them into a common representation, for example live movies and fixed samples? This is the question of data fusion. We proposed an approach to answer this question (Villoutreix et al., 2017), using a regression model between several sets of images containing various labels (e.g. DAPI or nuclei staining, protein staining, RNA, time stamps). Using the model learned, we could predict an integrative representation of multiple patterns of gene expression at high spatial and temporal resolution.

These examples illustrate that machine learning provides us with powerful methods and deep learning is particularly well suited for many of the challenges of microscopy in developmental biology. It should be noted that there is a large diversity of deep learning architectures (ways of defining a model to be learned) with many hyperparameters that define the characteristics of a model, which need to be tuned and can change the results.

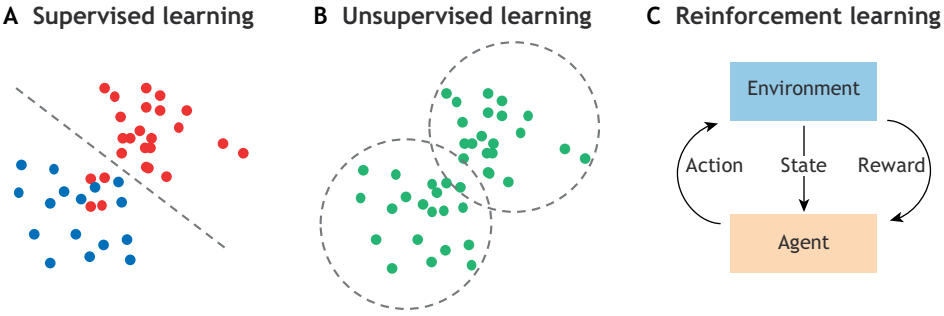

Fig. 2. Machine learning subdivision into supervised, unsupervised and reinforcement learning. (A) In the supervised setting, the points are labeled (in red and blue) and a decision criterion (dashed line) is learned using these labels. (B) In the unsupervised setting, the points are unlabeled (they are all green) and the clusters (represented by the dashed circles) are learned from the relationships between the points. (C) In the reinforcement learning setting, the agent (the computer) interacts with its environment to gain information through reward and to update its state accordingly.

Development (2021) 148, dev188474. doi:10.1242/dev.188474

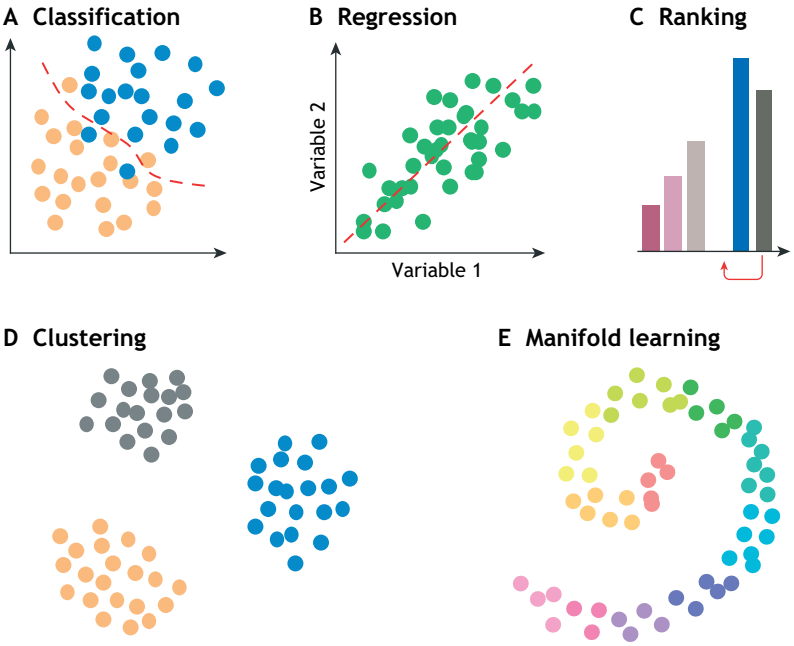

Fig. 3. Schematic of the common tasks of machine learning. (A) Classification aims to identify a classification criterion between two types of data points, represented here by the red dashed curve. (B) Regression aims to infer the relationship between two variables, represented by the red dashed straight line. (C) Ranking aims to order data according to their relationships; here, the rightmost bar is misplaced and should be ordered between the blue and gray bars. (D) Clustering aims to infer the class of each point, represented by each of the three colours, based on the relationships between them. (E) Manifold learning aims to infer the non-linear geometry of a point cloud, illustrated here by the gradient of colour along the spiral.

Additionally, supervised learning approaches rely heavily on manually labeled data, which can be costly and cumbersome to acquire. Training the dataset is crucial, because the quality of the results of any supervised learning algorithm depends on the quality of the training dataset. The size of the training dataset is also an important factor; if the model has too many parameters and is trained on a too small dataset, one risk is that the data could be overfit and therefore have poor generalization power. Machine learning approaches still need expert knowledge, both in biology and in computer science, to examine the results cautiously and to tune the models carefully.

Single-cell omics

Single-cell omics are a family of methods that take advantage of high-throughput sequencing techniques, such as single-cell RNA sequencing (Briggs et al., 2018; Wagner et al., 2018; Farrell et al., 2018) and Hi-C (Nagano et al., 2013), to measure the quantity of transcripts or the 3D conformation of DNA, respectively. Here, I show how machine learning can be used to study these large datasets.

Clustering

One of the main uses of single-cell transcriptomics in the past few years has been the reconstruction of differentiation trajectories in developing embryos. In particular, a series of papers have shown the progression of gene expression for every cell in a developing embryo (Briggs et al., 2018; Wagner et al., 2018; Farrell et al., 2018). The principal issue with this approach is that the single-cell data is obtained as a series of ‘snapshots’ at different developmental time points; to be able to measure the content of every cell in a tissue, the tissue has to be dissociated, thus stopping development. Therefore, differentiation dynamics are reconstructed from several embryos at different stages and then merged together. There are several challenges associated with this type of measurement. The first challenge comes from the high dimensionality of the data; here, a dimension refers to the measured quantity of transcript for one gene. In a typical single-cell RNA-sequencing experiment, there are about 20,000 dimensions. The second challenge is how to infer trajectories from biologically independent embryos. The introduction of non-linear, low-dimensional embedding techniques have been particularly successful in the study of high-dimensional data. These techniques aim to reduce the dimensionality of the data, while preserving its main features. The popular principal component analysis (PCA) is a linear way of reducing dimensionality, where the new representation is linearly related to the high-dimensional one. Non-linear dimensionality reduction techniques generalize this idea when the datasets have non-linear intrinsic structures. They are sometimes also called manifold learning techniques, the manifold being the geometric shape of the data points (Fig. 3E). One application of such techniques to transcriptomic data was the use of t-SNE (t-distributed stochastic neighbor embedding) by Dana Pe’er’s group in a framework called viSNE (Amir et al., 2013). More recently, UMAP (uniform manifold approximation and projection) has become more widely used, because of how well the topology of the data points is preserved when reducing the dimensionality. For example, a line in high dimension will stay a line in low dimension and similarly for a circle. The topological properties carry the qualitative relationships within the data (Becht et al., 2019). Usually, the mapping between high dimensions and low dimensions is inferred in a way that preserves the proximity relationships between the measured data points, enabling the visualization of clusters of similar cells, as well as the dynamics of how cells are diverging from each other, for example during development. The methods to explore these results are usually clustering methods and regression (Fig. 3B,D).

Trajectory inference

Several methods have been designed to reconstruct the trajectories of differentiating cells. When all the cells are undergoing the same process, they will appear to follow one trend in the low-dimensional embedding, which is what is used by pseudo-time inference methods (Trapnell et al., 2014; Saelens et al., 2019). This one-dimensional trend serves as an ordering of the cells, which can be one-to-one in relation to the actual time. When the shape of the manifold on which the points are lying is more complicated than just a line, which is the case for developing embryos, more sophisticated methods have been developed, relying on geometry inference methods for discrete point clouds (Farrell et al., 2018). Going

Development (2021) 148, dev188474. doi:10.1242/dev.188474

Box 1. What is deep learning?

Deep learning came to prominence in 2012, when its performance of image classification tasks was shown to be highly superior to any other approach (Krizhevsky et al., 2017 ; LeCun et al., 2015). We use elements of LeCun et al. (2015) to summarize this subfield of machine learning. Deep learning is a set of computational models that can learn representations of data with multiple levels of abstraction. They are called 'deep', because they are composed of a network of multiple processing layers, the depth of which enables representation of very complex functions. Each layer is composed of simple, but non-linear, units (the artificial neurons), that transform the representation at one level (starting with the raw data) into a representation in the next level in a more abstract way; together, they form a deep neural network. For a classification task (Fig. 3A), the higher levels will represent features that have a strong discriminative power, whereas the irrelevant variations will not be included. Many natural signals, like images, contain high-level features that can be represented by a composition of low-level features; this is the property that is successfully exploited by deep neural networks. The main challenge when considering multilayered network architectures is to find the weights associated with each subunit of the model, i.e. to learn a good representation of the data, millions of parameters have to be estimated. Gradient descent is the method used to optimize the accuracy of the resulting layer to the task at hand. It is coupled to a technique called backpropagation, which characterizes the joint variation of all the layers together when updating each of the weights within the network. Gradient descent and backpropagation are key to training deep neural networks.

further, some authors have attempted to fuse data from different sources to be able to infer correlation between variables that cannot be measured at the same time, or to use joint measurements for the embedding (Stuart and Satija, 2019; Liu et al., 2019 preprint). Overall these methods fall within the umbrella of non-linear dimensionality reduction methods, which are unsupervised techniques (Fig. 2B). The main limitation of non-linear dimensionality reduction methods comes from the difficulty of getting explicit mappings between the initial high-dimensional space and the final low-dimensional embedding; the new coordinate system is dependent on the data set used, making it therefore difficult to transfer to new situations.

Inference of spatial and temporal relationships

Finally, one recent interesting trend has been to devise computational approaches based on the mathematical framework of optimal transport to predict the spatial organization of cells from their transcriptomic profile (Schiebinger et al., 2019; Nitzan et al., 2019). The theory of optimal transport was originally developed by the French mathematician Gaspard Monge in 1781 to solve the problem of redistributing earth for the purpose of building fortifications with minimal work. Now, it is a geometric approach that aims at minimizing the distortion induced in mapping probability distributions to one another. In a study by Schiebinger and colleagues, differentiation trajectories within a large number of data points are inferred through the estimation with optimal transport of the coupling between consecutive time points (Schiebinger et al., 2019). Nitzan and colleagues make the hypothesis that the distance between cells in transcriptomic space is preserved in the physical space of a tissue, i.e. cells that are close in a tissue should have similar gene expression and vice versa (Nitzan et al., 2019). This principle, formalized as an optimal transport problem and applied to entire sets of cells, leads to accurate results, with data from zebrafish and fly development. It is an unsupervised learning method, in the same way as non-linear dimensionality reduction approaches, and is thus dependent on the dataset used for training so does not offer independent coordinate systems (Fig. 2B). Therefore, much care must be taken when generalizing to new situations. However, these two studies start to tackle some of the main challenges of single-cell transcriptomics, concerning its inability to measure spatial information or temporal information directly.

Growing an interdisciplinary community

Overall, the applications of machine learning in developmental biology are broad and the methods presented above are only a starting point. Now, I describe the conditions that could foster new interdisciplinary developments. The machine learning community is successful at developing methods that can be of broad use for data-intensive sciences. To grow an interdisciplinary community that would be able to fruitfully transfer knowledge between machine learning and developmental biology, we need to (1) establish well-curated databases (Allan et al., 2012; Pierce et al., 2019), (2) establish well-identified common tasks (Regev et al., 2017; Thul et al., 2017), (3) establish computational platforms to run algorithms (McQuin et al., 2018; Haase et al., 2020) and (4) improve computational literacy in biologist communities (Ouyang et al., 2019; Caicedo et al., 2019).

Machine learning methods are successful because they can take advantage of very large datasets. This requires infrastructure and the organization of research to advance science at the level of a community. The main need when developing machine learning models is labeled data used for training, where the relationship between the source and the target is precisely established. Published datasets can be used for many purposes, including being reused as training sets for machine learning. In order to take full advantage of the wealth of data that has been published – or is sitting siloed in labs, unused – we can use large storage infrastructures and data management tools, such as the platform OMERO for microscopy images. OMERO has been developed by the Open Microscopy Environment and is intended for accessing, processing and sharing scientific image data and metadata (Allan et al., 2012). In addition, to increase the speed of data dissemination we need to establish ways of citing and reusing data without going through the process of publishing peer-reviewed publications (Pierce et al., 2019).

Another way of organizing data at a large scale is to create common goals. The Human Cell Atlas and the Human Protein Atlas, which aim at providing reference maps for every human cell and every human protein within cells, respectively, are excellent examples of how to pool resources of various research teams (Regev et al., 2017; Thul et al., 2017). Researchers can share their data and benefit from the methods and insights stemming from the shared datasets. Similarly, competitions around one task are a good way of incentivizing computer scientists to develop the best algorithms, and can go a long way towards establishing standard problems and methods. One example is the 2018 Data Science Bowl challenge, which established the problem of segmentation in microscopy imaging and served as a source of data for various articles (Caicedo et al., 2019).

In addition to the question of how to store data, the second bottleneck is computational resources. As we have discussed, machine learning methods, and deep learning in particular, require specific hardware such as GPUs. The usual image processing software packages used in developmental biology are adapting to this growing need. For example, it is now possible to combine image processing software, such as FIJI or CellProfiler 3.0, with a GPU extension (McQuin et al., 2018; Haase et al., 2020).

Development (2021) 148, dev188474. doi:10.1242/dev.188474

Finally, it is important to acknowledge that designing a deep learning model is still a difficult task for someone with little experience in programming. Fortunately, accessible platforms, such as ZeroCostDL4Mic or ImJoy, can help bridge the gap between model users and developers (Ouyang et al., 2019; Von Chamier et al., 2020 preprint). Furthermore, computer science literacy initiatives are crucial to prevent a divide between computer scientists and biologists. NEUBIAS, the Network of European BioImage Analysists (http://eubias.org/NEUBIAS/training-schools/neubias-academy-home/), is an excellent example of such effort.

Perspectives

The promises of machine learning are high; however, there are several limitations that need to be mentioned. First, machine learning usually involves a 'black box'; when considering deep learning, there are millions of parameters that are trained to obtain a predictive model. These parameters are coupled in a non-linear way and it is therefore very difficult to understand how a deep learning model is actually making decisions. This is a major caveat, particularly when considering the potential applications to the medical field, where responsibility in decision making is crucial. Luckily, there are various ways to make a predictive model interpretable and open the black box (Gilpin et al., 2018). Importantly, this interpretability can even be used for biological discovery (Zaritsky et al., 2020 preprint). The second limitation of machine learning comes from the fact that the generalization power of these methods (i.e. their ability for extrapolation) is dependent on the training datasets. Indeed, the model can only extract the information of the dataset that has been used for training. There is some theoretical work that provides bounds to the generalization error but they are not always practical because they are not necessarily computable in a reasonable time. An example of this is the Rademacher complexity, which helps compute the representativeness of a dataset (how well an underlying true distribution is represented in the data) (Mohri et al., 2018). The practical way of characterizing generalization error consists of using cross-validation techniques (Friedman et al., 2009): the dataset is split into subparts, upon which the accuracy of the model can be evaluated. Cross-validation methods are, unfortunately, not sufficient in every situation, in particular when a training dataset is too small and leads to model overfitting or hallucinations. Moreover, when considering deep learning models, even though their high accuracy has been proved broadly, they are not immune to specific adversarial examples (Finlayson et al., 2019). For example, changing a few pixels of a chest X-ray can lead the classifier to miscategorize a healthy chest into a pneumothorax and vice versa. Adversarial examples are engineered for misclassification and expose the vulnerabilities of deep learning models. They can become a threat if a classifier is used routinely in a key real-world application. We also discussed the fact that non-linear dimensionality reductions methods extensively used in single-cell omics need to be used with caution as well, as their results are highly dependent on the datasets used. Finally, as mentioned previously, there is a need for expert knowledge to tune models accurately and examine their results, giving biologists and computer scientists important roles and responsibilities.

For developmental biologists, it is interesting to observe that machine learning has been largely inspired by research in computational neuroscience; indeed, deep learning has been built as an analogy to the circuitry of a brain. Developmental biology is a science that is concerned with the study of organized and functional systems emerging from individual units (cells). One can wonder if developmental biologists could benefit from implementing the principles of development into the algorithms of machine learning. One successful attempt is the so-called Compositional Pattern Producing Network (CPPN), which mimics pattern formation in development to infer any kind of function (Stanley, 2007). More precisely, this technique models the emergence of complex patterns in space (on 2D or 3D domains) by composing very simple functions (e.g. linear, quadratic or sinusoidal functions applied on the entire domain) until a complex pattern is obtained. The final pattern can be considered a composition of simple spatial patterns of gene expression, such as gradients. It has recently been used to automatically design reconfigurable organisms *in silico*, which have subsequently been generated *in vitro* (Kriegman et al., 2020). Cells and tissues as cognitive agents could serve as an inspiration and a model for autonomous computers and robots, evolving and developing in a complex environment.

In the not-too-distant future, we can expect that machine learning will be a part of many fields in the life sciences. Few studies have taken advantage of reinforcement learning in developmental biology and the intersection between machine learning and physical inference is only beginning (Gilpin, 2019). Finally, machine learning is a quickly changing field; new paradigms such as federated learning (Yang et al., 2019) or self-supervised learning (Jing and Tian, 2020) are emerging. Federated learning aims to take advantage of distributed data acquisition devices. In this paradigm, instead of updating machine learning models on a centralized data repository, the models are updated first locally on the data acquisition devices when new data is being acquired. The local models are then aggregated globally, without any data transfer, protecting data privacy and reducing the size of centralized data storage infrastructures (Yang et al., 2019). Self-supervised learning aims to overcome the need of labeled data for training deep learning models. In this approach, labels are generated automatically from large-scale unlabeled images or videos without any human intervention. Recent results show that the performance of self-supervised methods are comparable to supervised methods for object detection and segmentation tasks (Jing and Tian, 2020). Overall, progress in data management and machine learning model performance will likely lead to new breakthroughs in developmental biology.

Acknowledgements

I would like to thank the anonymous reviewers for insightful comments on the various versions of the manuscript.

Competing interests

The author declares no competing or financial interests.

Funding

This work has been funded by the Turing Center for Living Systems of Aix-Marseille Université.

References

Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W. J., Neves, C., Patterson, A. et al. (2012). OMERO: flexible, model-driven data management for experimental biology. *Nat. Methods* **9**, 245-253. doi:10.1038/nmeth.1896

Amir, E.-D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., Pe'er, D. et al. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545-552. doi:10.1038/nbt.2594

Azevedo, R. B. R., Lohaus, R., Braun, V., Gumbel, M., Umamaheshwar, M., Agapow, P.-M., Houthoofd, W., Platzer, U., Borgonie, G., Meinzer, H.-P. et al. (2005). The simplicity of metazoan cell lineages. *Nature* **433**, 152-156. doi:10.1038/nature03178

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F. and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38-44. doi:10.1038/nbt.4314

Belthangady, C. and Royer, L. A. (2019). Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat. Methods* **16**, 1215-1225. doi:10.1038/s41592-019-0458-z

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W. and Klein, A. M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780. doi:10.1126/science.aar5780

Caicedo, J. C., Goodman, A., Karhohs, K. W., Cimini, B. A., Ackerman, J., Haghighi, M., Heng, C. K., Becker, T., Doan, M., McQuin, C. et al. (2019). Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nat. Methods* **16**, 1247-1253. doi:10.1038/s41592-019-0612-7

Christiansen, E. M., Yang, S. J., Ando, D. M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O'Neil, A., Shah, K., Lee, A. K. et al. (2018). In silico labeling: predicting fluorescent labels in unlabeled images. *Cell* **173**, 792-803.e19. doi:10.1016/j.cell.2018.03.040

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A. and Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131. doi:10.1126/science.aar3131

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L. and Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science* **363**, 1287-1289. doi:10.1126/science.aaw4399

Friedman, J., Hastie, T. and Tibshirani, R. (2009). *The Elements of Statistical Learning*. New York, USA: Springer.

Gilpin, W. (2019). Cellular automata as convolutional neural networks. *Phys. Rev. E* **100**, 032402. doi:10.1103/PhysRevE.100.032402

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics, 80-89. doi:10.1109/DSAA.2018.00018

Haase, R., Royer, L. A., Steinbach, P., Schmidt, D., Dibrov, A., Schmidt, U., Weigert, M., Maghelli, N., Tomancak, P., Jug, F. et al. (2020). CLIJ: GPU-accelerated image processing for everyone. *Nat. Methods* **17**, 5-6. doi:10.1038/s41592-019-0650-1

Hollandi, R., Szkalisity, A., Toth, T., Tasnadi, E., Molnar, C., Mathe, B., Grexa, I., Molnar, J., Balind, A., Gorbe, M. et al. (2020). nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.* **10**, 453-458.e6. doi:10.1016/j.cels.2020.04.003

Jing, L. and Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* Early access, 1. doi:10.1109/TPAMI.2020.2992393

Khosravi, P., Kazemi, E., Zhan, Q., Malmsten, J. E., Toschi, M., Zisimopoulos, P., Sigaras, A., Lavery, S., Cooper, L. A. D., Hickman, C. et al. (2019). Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit. Med.* **2**, 21. doi:10.1038/s41746-019-0096-y

Kriegman, S., Blackiston, D., Levin, M. and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proc. Natl. Acad. Sci.* **117**, 1853-1859. doi:10.1073/pnas.1910837117

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84-90. doi:10.1145/3065386

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* **521**, 436-444. doi:10.1038/nature14539

Liu, T.-L., Upadhyayula, S., Milkie, D. E., Singh, V., Wang, K., Swinburne, I. A., Mosaliganti, K. R., Collins, Z. M., Hiscock, T. W., Shea, J. et al. (2018). Observing the cell in its native state: imaging subcellular dynamics in multicellular organisms. *Science* **360**, eaaq1392. doi:10.1126/science.aaq1392

Liu, J., Huang, Y., Singh, R., Vert, J. P. and Noble, W. S. (2019). Jointly embedding multiple single-cell omics measurements. *bioRxiv*. doi:10.1101/644310

McDole, K., Guignard, L., Amat, F., Berger, A., Malandain, G., Royer, L. A., Turaga, S. C., Branson, K. and Keller, P. J. (2018). In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. *Cell* **175**, 859-876.e33. doi:10.1016/j.cell.2018.09.031

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., Doan, M., Ding, L., Rafelski, S. M., Thirstrup, D. et al. (2018). CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol.* **16**, e2005970. doi:10.1371/journal.pbio.2005970

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M. and Van Valen, D. (2019). Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233-1246. doi:10.1038/s41592-019-0403-1

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018). *Foundations of Machine Learning*. Cambridge, USA: MIT Press.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A. and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64. doi:10.1038/nature12593

Nitzan, M., Karaiskos, N., Friedman, N. and Rajewsky, N. (2019). Gene expression cartography. *Nature* **576**, 132-137. doi:10.1038/s41586-019-1773-3

Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. and Johnson, G. R. (2018). Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917-920. doi:10.1038/s41592-018-0111-2

Ouyang, W., Aristov, A., Lelek, M., Hao, X. and Zimmer, C. (2018). Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnol.* **36**, 460-468. doi:10.1038/nbt.4106

Ouyang, W., Mueller, F., Hjelmare, M., Lundberg, E. and Zimmer, C. (2019). ImJoy: an open-source computational platform for the deep learning era. *Nat. Methods* **16**, 1199-1200. doi:10.1038/s41592-019-0627-0

Pierce, H. H., Dev, A., Statham, E. and Bierer, B. E. (2019). Credit data generators for data reuse. *Nature* **570**, 30-32. doi:10.1038/d41586-019-01715-4

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M. et al. (2017). Science forum: the human cell atlas. *eLife* **6**, e27041. doi:10.7554/eLife.27041

Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547-554. doi:10.1038/s41587-019-0071-9

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P. et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928-943.e22. doi:10.1016/j.cell.2019.01.006

Schmidt, U., Weigert, M., Broaddus, C. and Myers, G. (2018). Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention — MICCAI 2018* (ed. A. Frangi, J. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger). Lecture Notes in Computer Science, vol 11071. Springer, Cham. doi:10.1007/978-3-030-00934-2_30

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489. doi:10.1038/nature16961

Stanley, K. O. (2007). Compositional pattern producing networks: a novel abstraction of development. *Genet. Program. Evol. Mach.* **8**, 131-162. doi:10.1007/s10710-007-9028-8

Stringer, C., Wang, T., Michaelos, M. and Pachitariu, M. (2020). Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods* **14**, 1-7. doi:10.1038/s41592-020-01018-x

Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257-272. doi:10.1038/s41576-019-0093-7

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M. et al. (2017). A subcellular map of the human proteome. *Science* **356**, eaal3321. doi:10.1126/science.aal3321

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., Rinn, J. L. et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381-386. doi:10.1038/nbt.2859

Villoutreix, P., Andén, J., Lim, B., Lu, H., Kevrekidis, I. G., Singer, A. and Shvartsman, S. Y. (2017). Synthesizing developmental trajectories. *PLoS Comput. Biol.* **13**, e1005742. doi:10.1371/journal.pcbi.1005742

Von Chamier, L., Jukkala, J., Spahn, C., Lerche, M., Hernández-Pérez, S., Mattila, P., Karinou, E., Holden, S., Solak, A. C., Krull, A. et al. (2020). ZeroCostDL4Mic: an open platform to simplify access and use of deep-learning in microscopy. *bioRxiv*. doi:10.1101/2020.03.20.000133

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G. and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987. doi:10.1126/science.aar4362

Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., Schmidt, D., Broaddus, C., Culley, S. et al. (2018). Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nat. Methods* **15**, 1090-1097. doi:10.1038/s41592-018-0216-7

Weigert, M., Schmidt, U., Haase, R., Sugawara, K. and Myers, G. (2020). Star-convex polyhedra for 3d object detection and segmentation in microscopy. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3655-3662. doi:10.1109/WACV45572.2020.9093435

Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019). Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **10**, 1-19.

Zaritsky, A., Jamieson, A. R., Welf, E. S., Nevarez, A., Cillay, J., Eskiocak, U., Cantarel, B. L. and Danuser, G. (2020). Interpretable deep learning of label-free live cell images uncovers functional hallmarks of highly-metastatic melanoma. *bioRxiv*. doi:10.1101/2020.05.15.096628