What do machines see? Utilizing artificial intelligence to explore cell biology

Matt De Vries and **Chris Bakal** (The Institute of Cancer Research, UK)

With our ability to take and quantify numerous complex images of cells and cell populations, the ability to paint an accurate picture of the underlying data has never been more valuable. Deferring from the contemporary classics in data visualization to methods that exploit advances in artificial intelligence is an essential step in understanding high-throughput, three-dimensional microscopy data. This feature article discusses how generating or simulating representative cells that may not exist in the data set, yet summarize the underlying distribution, allows researchers to effectively and efficiently analyse cellular morpho-dynamics. Furthermore, learning from these artificial intelligence-based techniques allows us to 'see what the machine is seeing' in a step towards unpacking the chaos of cell biology to understand the very fundamentals of living organisms.

Visualizing phenotypic occurrences in high-throughput experiments

Cell biology is founded on the principle of 'seeing is believing'. Due to this natural essence, examining the images themselves has historically been involved in the presentation of microscopy-based data. With developments in computer vision, we can now extract a plethora of information from images. However, as the volume and depth of images that can now be generated in single experiments continues to grow, the availability of techniques to effectively analyse phenotypic differences decreases.

With our ability to take and quantify more images, the ability to paint a clear and accurate picture of the underlying data has never been more valuable. Visualization delivers data in the most efficient way possible. As an essential step in the scientific process, data visualization takes raw data, models it through varying mathematical methods and delivers it so that conclusions on trends, patterns and outliers within large sets are more easily comprehensible.

Visualizing image data

Contrary to intuition, data visualization of image data is non-trivial. Raw image data does more to ask questions than guide scientists through the full 'picture' (Figure 1a). This becomes even more apparent when adding volumetric and time dimensions to make 3D movies of objects or multiple objects. In particular, humans are more so concerned with smaller components of larger structures in images. These may include facial expressions in pictures of several faces, analysing human gait as a biometric in low-resolution CCTV footage, and subcellular structures in cell communities (Figure 1b). These complex characteristics of imaging data require visualization techniques beyond the contemporary classics.

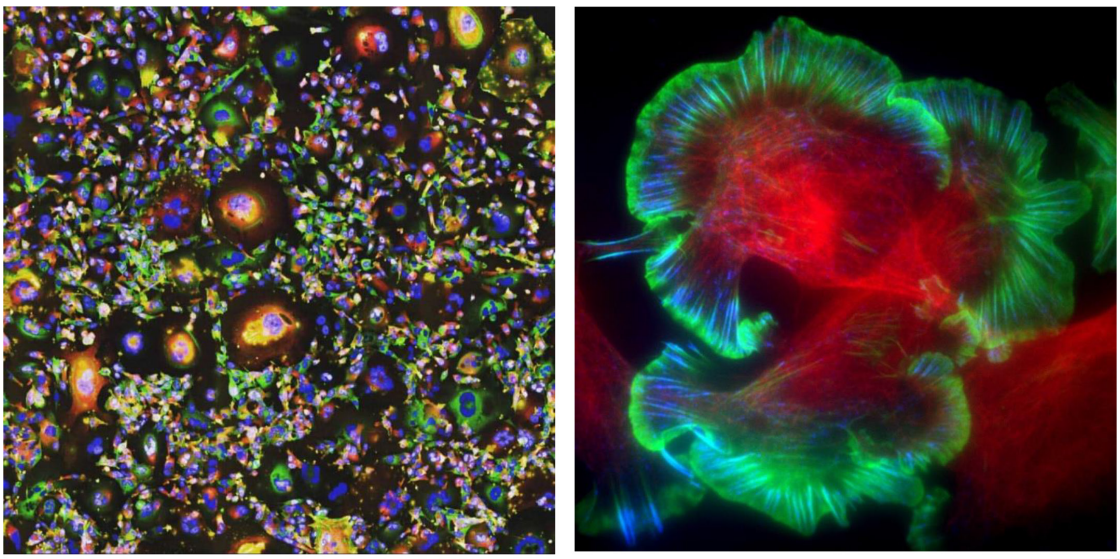
In addition to new abilities to image more cells at higher resolutions, computer vision technologies now allow us to quantify different aspects of the image. Such methods turn images into numbers and have important consequences on how data visualization is performed. Quantifying and presenting fundamental features of cell morphology were pioneered in fixed-cell assays and extended to live-cell dynamics. A common theme has been to use these raw features to build computational models to study the distribution and group phenotypes.

We may then visualize these groups by generating representative cells for each. This has been done through classical machine learning methods as well as deep learning. An additional dimension to the data increases complexity exponentially. In fact, we learn from Pólya's recurrence Theorem[i] that it is very easy to get 'lost' in 3D. We may extend some of the basic

[i]Pólya's recurrence theorem states: a simple random walk on a d-dimensional lattice is recurrent for d = 1,2 and transient for d >2. This means that a random walk done in 2D ensures recurrence with 100% probability. A random walk done in 3D only ensures recurrence with roughly 34% probability. This has

 © The Authors. Published by Portland Press Limited under the Creative Commons Attribution License 4.0 (CC BY-NC-ND)

Downloaded from http://portlandpress.com/biochemist/article-pdf/43/5/48/922616/bio_2021_168.pdf by Institute of Cancer Research user on 22 July 2022



(a) (b)

Figure 1. Complexity in raw microscopic images. (**a**) Isogenic triple negative breast cancer cells showing a wide range of shapes and phenotypes. A clear example of cancer heterogeneity and how their analysis may prove challenging even to experts. Image taken by Patricia Pascal-Vargas. (**b**) Melanoma cell spreading on collagen matrix. These cells are master shape-shifters, which rapidly adapt to new environments. Here we see the subcellular structures carry an abundance of information with microbial adhesions (cellular grips) represented in blue, showing their morphological use in meeting the physical demands of new tissue. Image taken by Oliver Inge and Chris Bakal.

features commonly used in 2D assays to 3D; however, they may not be exhaustive. On top of predefined features to represent data, we are able to learn abstract features through deep learning. These are common problems in computer vision tasks where classes are relatively rudimentary. Modelling and representing data that follow distributions far more convoluted allow researchers to be guided by technological advances.

Deep learning and visualization

Deep learning techniques have made an inherent impact on computer vision tasks and – due to the central role of imaging – biological sciences. Previously, these techniques have been used as 'black-box' methods. However, in tasks where the question of 'why' matters most, substantial explainability needs to be considered. These algorithms aim to model distributions in and of multiple images, obtain abstract features and simplify complex correlations. Importantly, these algorithms are capable of exploring intricate data in a much more efficient and effective way than humans. Ultimately, scientists want to know why these methods work so well and learn from them. What are these algorithms 'seeing'?

Simulating and representing cell populations

Simulation allows us to visualize data that follow specific rules and directly change features to understand their impact. Early simulations aimed at exploring fundamental biological processes as well as the hallmarks of cancer. Using these models, we can visualize and even digitally manipulate cell behaviour. Understanding large datasets containing thousands (or millions) of heterogeneous cells with multiple phenotypes may be done through exploring a few representative or exemplar cells. These cells may not actually exist in the dataset but are characteristic or provide more contextual information (Figure 2). One may view these representative cells of complex image datasets similar to summary statistics of more simple tabular datasets.

We may also make use of explainable artificial intelligence in exploring why algorithms have grouped specific data. Methods known as class activation mapping try to explore what is going on when a machine learning algorithm makes particular decisions (Figure 3). For example, certain neurons in a deep neural network will activate or 'show attention' when looking at cells that it categorizes as similar. Through mathematical operations

more pure mathematical value than that of qualitative visualization, but allows insight into how dimension greatly affects complexity.

Downloaded from http://portlandpress.com/biochemist/article-pdf/43/5/48/922616/bio_2021_168.pdf by Institute of Cancer Research user on 22 July 2022

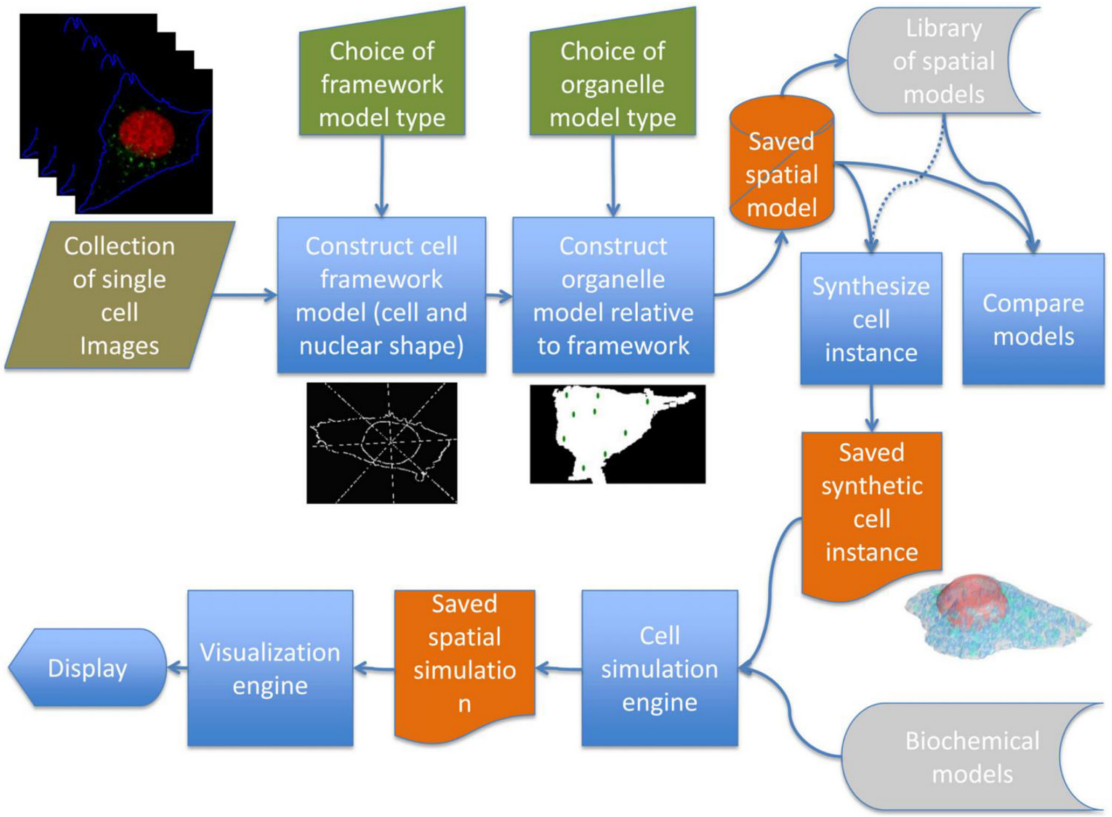


Figure 2. A diagram of generating and visualizing representative cells. After collecting a large database of single cells, morphological profiles for each cell are calculated. These morphological profiles include shape descriptors as well as subcellular structures. Varying mathematical techniques are used to model these profiles and find representative cells for certain classes. These representative cells are then simulated through the synthesis of multiple subcellular components and finally visualized using a visualization engine such as Napari or Fiji. [Source: Murphy, 2016, Methods, DOI: 10.1016/j.ymeth.2015.10.011].

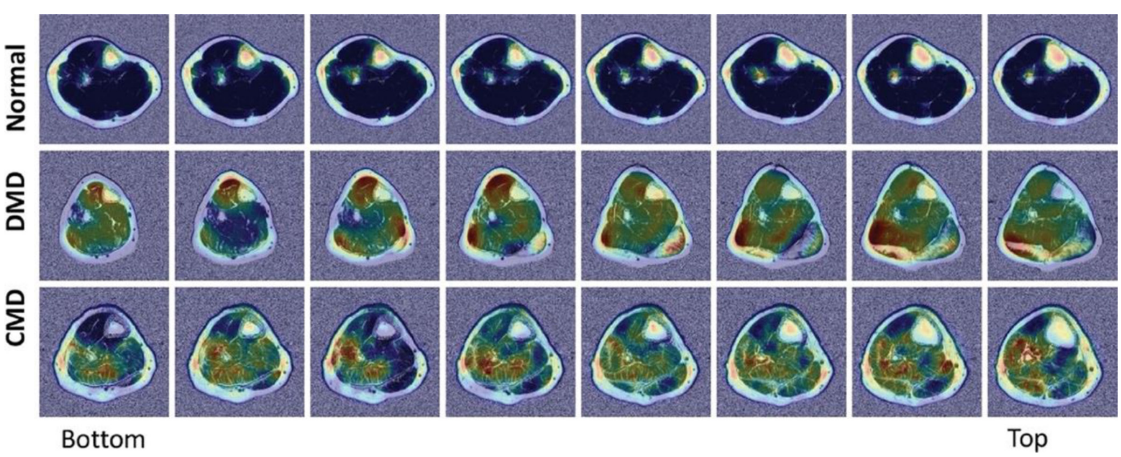


Figure 3. Class activation mapping 'unboxing' the black box that is deep learning. Here, we see that class activation mapping on tasks of muscular dystrophy classification in MRI is able to explain areas of most importance when a model classifies images as normal, Duchenne muscular dystrophy and congenital muscular dystrophies. These areas of most importance were shown to correlate well with biological interpretation thus verifying the model. [Source: Cai et al., 2019, Pattern Recognition, DOI: 10.1016/j.patcog.2018.08.012].

Downloaded from http://portlandpress.com/biochemist/article-pdf/43/5/48/922616/bio_2021_168.pdf by Institute of Cancer Research user on 22 July 2022

 © The Authors. Published by Portland Press Limited under the Creative Commons Attribution License 4.0 (CC BY-NC-ND)

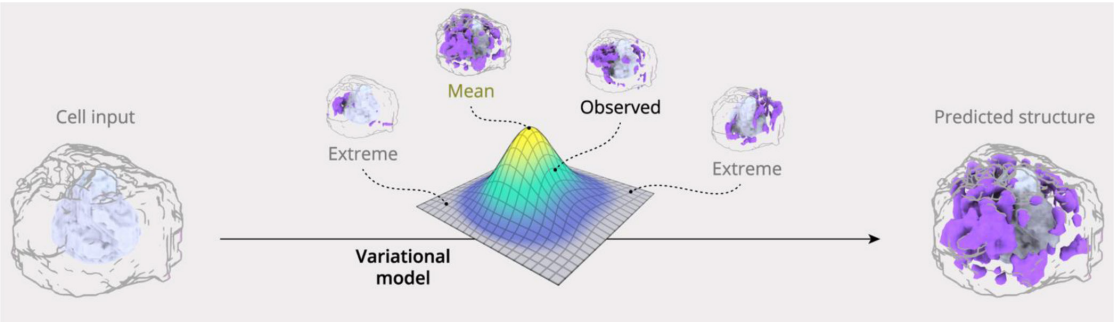


Figure 4. Probabilistic modelling allows the generation of a probabilistic representation that presents locations and morphologies of subcellular structures. This may then be incorporated in user-friendly graphic user interfaces to look deeper into cells through realistic simulations of cells' *in vivo* contexts [Source: https://www.allencell.org/3d-probabilistic-modeling.html].

known as convolution and pooling, we can project that 'attention' on the original data to essentially see what the machine 'sees'. Through this, we are also able to test the model by trying to trick it. Why does the model determine particular cells are of that phenotype, and if we warp that cell through various transformations, when will the model cease to recognize the phenotype?

Generating exemplar cells that complex computational models declare as being representative of certain phenotypes and 'real' (although they may not exist) is often done through deep learning techniques. Exploring and generating specific components of cells and how they work together in an integrated fashion allows us to truly understand the full context of cell biology. By training algorithms to recognize subcellular structures that even experts can't see, we may be better equipped to label and visualize them more clearly. Combining this with probabilistic modelling, which captures variations in cell morphologies and locations for all components of interest, cells may be simulated and visualized in a much richer sense than simply seeing them through the microscope (Figure 4).

Conclusion, challenges and future aspects

Up until recently, what we view as typical cells (drawings shown in biology textbooks) are a great starting point. However, in reality, there is most probably no cell that has ever looked like that. Visualizing what is going on in cells and cell communities has commonly been done through viewing raw microscopy images. This becomes exponentially more difficult when the dimension and throughput increases. Visualization techniques beyond the classic need to be explored to understand underlying cell processes, how each subcellular component works together and capture the heterogeneity found in cell populations.

Early attempts to quantify morphological features and present distinct classes have paved the way for more complex generation and simulation of representative cells through modern artificial intelligence-based techniques. This, along with the increasing computational power of today, has allowed researchers to create visualizations that give much more information than previously dreamed. We can now explore and change specific components and analyse how they affect the environment as a whole. We know that cells react differently to different environments and adding these kinds of features into simulations could extend our knowledge of the *in vivo* context.

Although we have come a long way, we are still limited by the technology of our time. These computational models are heavily reliant on the amount of data on which they are trained. Efforts towards high-throughput imaging will significantly increase the accuracy and effectiveness of these models. High-resolution volumetric time-series data inherently takes up significant amounts of storage, often more than that available on modern processing units, which help speed up computation. If we cannot fit the data and the model on these units, we cannot achieve results fast. Hardware companies are continuously pushing the boundaries and what was thought impossible a decade ago is already possible today.

Simulating complex environments accurately and efficiently will prove vital in understanding the complexity and unpacking the chaos of cell biology. Moreover, seeing what the machine sees is an excellent step in utilizing the ever-growing technology to understand the very fundamentals of living organisms. ■

Downloaded from http://portlandpress.com/biochemist/article-pdf/43/5/48/922616/bio_2021_168.pdf by Institute of Cancer Research user on 22 July 2022

Further reading

- Bakal, C., Aach, J., Church, G. and Perrimon, N. (2007) Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science* **316**, 1753–1756. DOI: 10.1126/science.1140324
- Sommer, C. and Gerlich, D.W. (2013) Machine learning in cell biology - teaching computers to recognize phenotypes. *J. Cell Sci.* **126**, 5529–5539. DOI: 10.1242/jcs.123604
- Murphy, R.F. (2016) Building cell models and simulations from microscope images. *Methods.* **96**, 33–39. DOI: 10.1016/j.ymeth.2015.10.011
- Schubert, P.J., Dorkenwald, S., Januszewski, M. et al. (2019) Learning cellular morphology with neural networks. *Nat. Commun*. **10**, 2736. DOI: 10.1038/s41467-019-10836-3
- Yao, K., Rochman, N.D. and Sun, S.X. (2019) Cell type classification and unsupervised morphological phenotyping from low-resolution images using deep learning. *Sci. Rep*. **9**, 13467. DOI: 10.1038/s41598-019-50010-9
- Alizadeh, E., Castle, J., Quirk, A. et al. (2020) Cellular morphological features are predictive markers of cancer cell state. *Comput. Biol. Med*. **126**, 104044. DOI: 10.1016/j.compbiomed.2020.104044
- Johnson, G.R., Donovan-Maiye, R.M. and Maleckar, M.M. (2017) Building a 3d integrated cell. *bioRxiv.* DOI: 10.1101/238378
- Wu, P.-H., Gilkes, D.M. Phillip, J.M. et al. (2020) Single-cell morphology encodes metastatic potential. *Sci. Adv*. **6**, DOI: 10.1126/sciadv.aaw6938
- Xue, Y., Wang, J., Ren, K. and Ji, J. (2021) Deep mining of subtle differences in cell morphology via deep learning. *Adv. Theory Simul*. **4**, 2000172. DOI: https://doi.org/10.1002/adts.202000172

Matt De Vries is a PhD candidate at the Institute of Cancer Research in London under the supervision of Professor Chris Bakal. His research is focussed on AI techniques to analyse complex imaging data to better understand the heterogeneity in cancer cell biology and the driving forces of metastasis.

Chris Bakal is the Professor of Cancer Morphodynamics at the Institute of Cancer Research in London, UK, where he leads the Dynamical Cell Systems Laboratory. His team aims to understand how signalling networks regulate cancer cell shape. Chris was born in Calgary, Canada. He received his BSc in Biochemistry from the University of British Columbia, and his PhD in Medical Biophysics from the University of Toronto. Chris' postdoctoral work was performed in Department of Genetics at Harvard Medical School, and the Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT). In 2007, Chris was named as one of the most promising postdoctoral fellows or junior faculty members at Harvard Medical School by the Dorsett L. Spurgeon award. After being awarded a Wellcome Trust Career Development Fellowship, Chris established his laboratory at the Institute of Cancer Research in London in 2009. In 2015 he was awarded the prestigious Cancer Research UK Future Leaders Prize.Outside of science Chris is competitive track cyclist, a former national-level runner, and a former world-ranked downhill ski racer. Chris has run a mile in just over 4 minutes, and aims to compete in the Ironman next year.

Downloaded from http://portlandpress.com/biochemist/article-pdf/43/5/48/922616/bio_2021_168.pdf by Institute of Cancer Research user on 22 July 2022

 October 2021 © The Authors. Published by Portland Press Limited under the Creative Commons Attribution License 4.0 (CC BY-NC-ND)