Voices

Which image-based phenotypes are most promising for using AI to understand cellular functions and why?

Emma Lundberg
Science for Life Laboratory, Royal Institute of Technology (KTH)

Finding rare phenotypes

Developments in microscopy and visualization techniques have long pushed the boundaries of our understanding of cells. With powerful machine-learning techniques, we're now facing a paradigm shift in image analysis for cell biology. Conventionally, image-based phenotypes have been described using a selected set of features we can easily intuit, such as shape or marker intensity. In the deep-learning era, neural network models can automatically learn representations of cellular phenotypes. These deep-learning features are data-driven, scalable, and sensitive to the representation of sophisticated or hidden patterns in the images. This comes with possibilities like continuous modeling of cell phenotypes and the discovery of rare or new phenotypes by deep anomaly detection.

The challenge will be to translate the unexpected phenotypes unraveled with machine learning into deeper biological understanding. Here, visual explanation techniques such as grad-CAM can help. But beyond the development of interpretable machine-learning models, I also believe that the imaging experiments need to be done differently. With an iterative data-driven experimental design, I can envision imaging schemes where labels of cellular structures and states are predicted in real time at the microscope, analyzed, and coupled to deeper imaging of the key phenotypes (and iterate again) until the needed quality and completeness of the dataset and analysis is reached.

Machine learning will enable integration of image data with multi-omics data to provide a dynamic and complex systems framework for understanding cells, particularly for phenotypes that are rare, transient, or subtle. Together with the establishment of public cell image repositories, one can envision microscopy reaching the point where re-analysis and meta-analysis across image datasets can help us answer previously intractable biological questions.

Jan Funke
HHMI Janelia Research Campus

The non-obvious ones

Of the many variables that describe the internal state of a cell, any imaging technique allows us to infer only a small subset through indirect measurements like fluorescent probes or changes in electron density. In general, there is no reason to believe that the relationship between those variables is simple, i.e., easy to understand by humans. In this context, data driven machine learning is promising and likely necessary to model non-linear correlations between phenotype and functional properties.

But here is the dilemma: we replace one complex system (imaging to indirectly measure inter-related variables) with another one (a highly non-linear computational model). As we celebrate the possibilities of machine learning in the life sciences, we should take note that without understanding computational models, we are limited to the automation of tasks that we can easily verify. For example, we can train a machine-learning algorithm to segment cell organelles, which allows us to analyze datasets that are otherwise prohibitively large. That is a big contribution in itself, but the looming question is: Can we use machine learning to tell us about relationships that we did not know of before?

The most promising phenotypes are therefore those for which we can develop "in-silico" probes that predict function from structure in non-obvious ways. Crucially, those models need to be interpretable. This is necessary to gain trust in the presence of dataset specific confounders. What is more, interpretable methods will teach us about the

learnt correlations and further our understanding of the relation between structure and function.

Chris Bakal
Institute of Cancer Research, UK

Cell shape predicts cell state

Every day, pathologists perform sophisticated diagnoses by visually examining tissue sections. Using only a microscope, a pathologist infers the existence of unseeable states, such as cancer type, aggressiveness, and even mutational status. How pathologists do this is not always clear. But gaining insight into how humans make decisions based on observable phenotypes is key to developing Artificial Intelligence based methods capable of automated pathology, or that are used as research tools.

The ability of humans to infer complex unseen states is because by (very) deep learning, the brain trains "classifiers" that use both images and prior data (genetics, biochemistry) as inputs. Essentially, the brain builds relationships between observable phenotypes, and what it has previously learned drives those phenotypes. For example, that the presence of a mutation correlates with an observable feature.

We argue cell shape is a key feature in classifiers which can predict cell states from images. Because cell shape correlates strongly with, and is sometimes even causal to, many aspects of cells and of proteins, we can often reduce high-dimensional phenotypes to cell shape alone. Toward reverse engineering what our brains do so well, we use machine-learning based methods to generate "Quantitative Morphological Signatures" that describe the shape of individual cells. Through integration of these signatures with other datasets they become predictive of gene function, protein localization, and protein-protein interaction. We believe cell shape is an informative and easy to measure feature that can even potentially replace technologies such as transcriptomics or proteomics in many types of analyses.

Virginie Uhlmann
EMBL-EBI

Embracing complexity

The wealth of complex features imaging data hold offers the ideal ground to explore machine learning's potential beyond sophisticated pattern recognition. Recent developments in (bio)image analysis provide us with efficient tools to quantify texture, shape, abundance, and localization of cellular components, and to characterize how any of these features evolve in time. The dynamics and interdependencies of the many image-based phenotypes hold more information than each of them separately. AI can provide tools to integrate this massive amount of quantified visual features into interpretable biomechanical models, thus bringing theoretical and experimental approaches to biology closer than they have ever been.

Modern machine learning indeed plays a pivotal role in the development of first-principle mathematical models. It enables using complex experimental datasets to solve elaborate dynamical models of physical phenomenon by efficiently exploring parameter spaces that would be daunting for classical optimization methods. It also goes beyond, enabling the use of data to discover partial differential equations. This is the direction explored by physics-informed neural networks, which learn solutions of supervised learning tasks that respect a prescribed set of physical laws described by general nonlinear partial differential equations.

We will not unravel the mechanisms underlying life's complexity using machine learning or mathematical modeling alone. However, combined into a virtuous circle of data mining, hypothesis formulation, and model validation, these complementary approaches will grant us the ability to exploit the full potential of bioimages, unify visual information with omics data in multi-modal models, and push the boundaries of our understanding of cellular function and dynamics.

Daniel Gerlich
Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA)

Collaboration is key

Combining automated microscopy with image analysis by computer vision and machine learning has paved the way for cell biological experiments at very high throughput. To investigate specific cell biological processes by automated imaging, appropriate fluorescence markers have to be identified and complex cell morphology feature profiles have to be converted into a phenotype score. Implementing sensitive and robust imaging-based assays remains challenging, as it requires thorough understanding of the underlying biology, microscopy, and computational image analysis procedures.

Developing a high-content screening assay should consider how phenotype assessment by humans differs from that by computer algorithms. Phenotype classification by machine learning typically relies entirely on the image content, whereas humans also take rich biological and technical background knowledge into account when interpreting images. For example, when humans observe changes in texture granularity or edge sharpness, they consider not only the effect of the cellular perturbations but also knowledge about how microscope focus inaccuracies or illumination irregularities affect cell morphology. Developing a sensitive and robust assay based on computational procedures thus requires careful adjustments of the cell biological experiment, fluorescence markers, and image analysis strategy to focus the phenotype score entirely on image contents and to eliminate indirect experimental side effects as much as possible. Efficient communication and close collaboration between cell biologists and computer scientists is key for successful high-content screening.

Thomas Walter
Mines ParisTech, Institut Curie

Beyond the human eye

Computer Vision has been the workhorse for computational phenotyping for more than 20 years. Application of single cell classification to High Content Screening data has allowed us to systematically study subcellular protein localization and functional phenotypes. Today, Deep Learning has revolutionized Computer Vision, reaching unprecedented performance in image classification and segmentation. The power of Deep Learning resides in learning very complex relationships between input images and output variables without relying on manually defined image features, provided there is sufficient ground truth data for training the models. The by far most interesting application case arises when the ground truth does not rely on manual annotation. For instance, we can train networks to predict fluorescent labels from transmitted-light microscopy, thereby allowing for comprehensive analyses covering many aspects of cellular phenotypes. The ground truth can also result from the individual developmental trjectory of each cell in order to identify the first morphological cues for fate decision. Yet another interesting approach is to predict molecular features (typically gene expression signatures) from image data, an approach that is currently being investigated in digital pathology. Altogether, Deep Learning has allowed us to go beyond automating annotations by making predictions that are not obvious to the human eye. This will provide us with exciting new hypotheses on the origin and consequences of morphological phenotypes

Anne Carpenter
Broad Institute of Harvard and MIT

Measure all the things!

Our laboratory often discovers morphological changes in cells that are responding to a drug treatment, a genetic perturbation, or a disease state. We do this by measuring as many features of the cell as we can (size, shape, stain intensity and texture, and so on), in a strategy called image-based profiling. Once we find a change, the first questions a biologist asks are "What is the change? What does it mean?" Answering these questions can unveil the underlying mechanisms of the drug, gene, or disease. So, it is great when you can get an answer. But I also want to encourage researchers to become more comfortable with those cases where it is *not* feasible to precisely pinpoint the morphological change, at least not in a way that meshes nicely with our existing biological knowledge of structures in the cell. Deep learning, with its lack of obviously interpretable features, has certainly pushed us into this realm. If proper experimental design and careful machine-learning controls tell us that a morphological change is real, we

shouldn't worry overly much about its biological interpretability, we should embrace just *using* the pattern for discovery—for example, identifying a drug that reverses a discovered disease phenotype. So, my short answer to "which phenotypes?" is "any and all of them!"
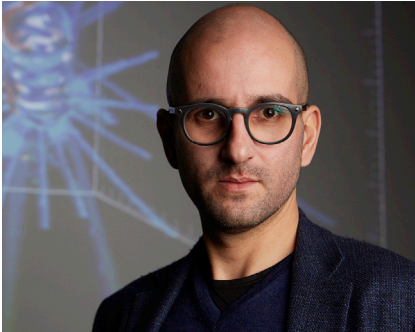
Luis Pedro Coelho
Fudan University

Image BLAST

When working with sequencing data, computational approaches can be classified into either reference-based (e.g., mapping reads to previously sequenced genomes or using BLAST) or *de novo* (e.g., building metagenome-assembled genomes from scratch). In the world of imaging, however, computational analyses are almost exclusively *de novo* and integrating the results with prior knowledge is performed by the human researcher. In fact, until recently, most image analysis methods had to be tweaked for each dataset, each imaging modality, often each individual molecular probe and microscope combination.

Recently however, mostly due to the advent of deep learning, more robust methods have emerged which can be applied even across different modalities. These are convenient but are still only addressing tasks such as segmentation or image classification. I think that the next big step is methods can exploit background information that is present in databases. This can include image databases, such as using previously acquired images to find similar phenotypes (so called "image BLAST," a concept that is not novel, but is still far from being fully realized), but also other modalities and predictions. For example, analyzing spatial transcriptomics data by using bulk-sequenced data and predicted gene networks as a reference. Whatever form it takes, it will require creativity: deep learning provides us with the engine, but designing the machines is still to be done.