APPLICATION OF MACHINE LEARNING TO CELLULAR AGRICULTURE

Zachary Cosenza
University of California
Davis
zacosenza@ucdavis.edu

Michael Todhunter
Beckman Research Institute
City of Hope
todhunter@todhunter.dev

ABSTRACT

In this short document, produced on behalf of *New Harvest* for the purpose of helping *Google* understand and explore possible collaborations, we highlight the potential use-cases of machine learning (ML) in the field of cellular agriculture (CA) production. This work is organized by classes of problems in cellular agriculture (media design, regulatory network discovery, phenotype analysis for example) where ML or computational techniques have been or might be useful. Because of the novelty of CA, few case studies exist in the field itself. Therefore, this work will draw upon adjacent fields in biotechnology, bioinformatics, cancer research etc.

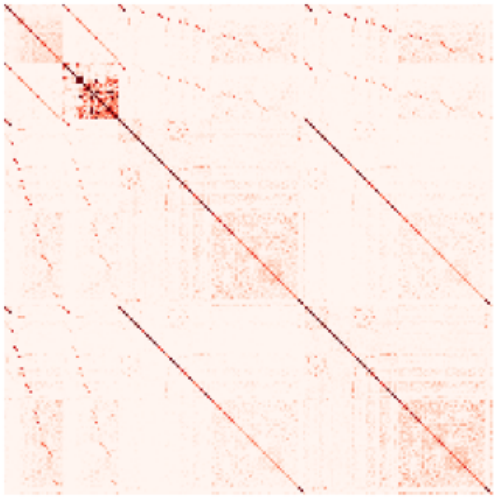
1 Introduction

CA (or cultured meat) is the production of animal products in vitro rather than through traditional animal husbandry [1]. This is done for a variety of reasons, primarily (i) reduction of environmental degradation of agriculture and (ii) reduction of animal suffering. While the processes to accomplish this task are varied and largely untested, typically animal cells are proliferated in large numbers then matured into tissues and final products to mimic their traditional structures [2]. While there are now many companies working on such products (`https://vegfaqs.com/lab-grown-meat-companies/`), many economic and technical issues must still be addressed [3] and should be carefully considered in the context of funding research. The focus of this work is on the applications of ML to CA to address some of these challenges, so we will focus the rest of this work on these applications including **media design, gene network discovery, multi-omics analysis, and phenotype analysis**. This is not meant to be an exhaustive list of potential applications of ML and related methods to CA, but document to demonstrate useful collaborations that are possible in the fields that we are familiar with.

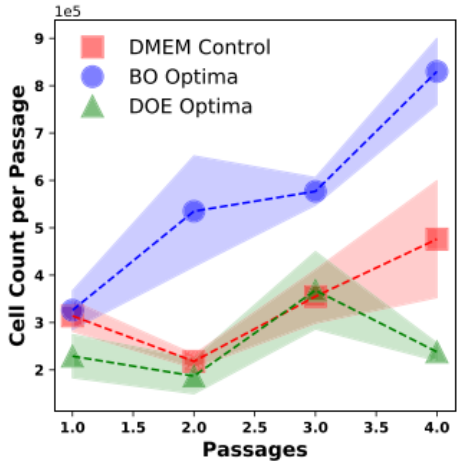
2 Active Learning and Media Design

The largest contributor to the cost of CA is the media that provides the cells with metabolites and growth factors so that the cells can grow (proliferate) and form meat-like structures (differentiation) [4]. Mammalian cells require complex mixtures of energy sources, amino acids, vitamins, salts, metals, osmolarity and shear force regulators, growth factors and other proteins in order to survive. For an excellent review of media see [5]. DMEM, a common growth media used in the laboratory setting, is comprised of > 30 components not including the numerous undefined components of the fetal bovine serum (FBS) typically paired with DMEM. To further complicate matters, the effects and interactions between all of these components is nonlinear and cell-type dependent.

One area where ML can alleviate the enormous expensive of developing and optimizing media with many components for CA applications is active learning (AL). AL is the array of techniques whereby data is selected to improve an underlying model or metric of interest in the most efficient way possible [6]. The general workflow for active learning processes is (i) collect some initial data, (ii) train a model, and (iii) use the model to predict the next best set of data to collect, (iv) repeat. This is similar to design-of-experiments (DOE) where experiments are designed in a single shot in order to, for example, learn the linear effects of each component (factorial design) or maximize signal-to-noise



(a) Covariance matrix of BO data.



(b) Long-term growth of optimal media.

Figure 1: AL example from [13] optimizing a 14 dimensional muscle cell (C2C12) media. Goal was to induce growth up to four passages of growth. BO and DOE Optima refer to media designed using a Bayesian optimization and traditional DOE method respectively. DMEM control refers to a typical medium used in laboratories to grow mammalian muscle cells. (a) Shows the covariance matrix generated by the multi-information source experimental campaign. Notice the four distinct diagonal blocks, which correspond to the four methods the authors used to measure cell growth. (b) Shows the Bayesian method designed experiments that resulted in far superior performance relative to a baseline optimizer (DOE) and commercial product (DMEM).

ratio. In [7] cell yield and medium cost was optimized using a genetic algorithm and local random search framework. A regularized radial basis function was used to model 30 metabolite concentrations to a yield / cost ratio of C2C12 muscle cells and locate optimal concentrations. Similar work was done by [8, 9] for *Lactococcus lactis* fermentation and [10] for *Synechococcus* cyanobacteria biomass yield and reaction conversion also using genetic algorithms. CA is particularly in need of good AL methods and further studies in AL-developed media because (i) the need for cost reduction and (ii) the biomass of cells needed are far greater than in typical biotech / pharmaceutical industry so media and process optimization will need to mimic lower margin industries like oil and gas or foods rather than biotechnology (where media is often a negligible contributor to cost).

Data-fusion and multi-information models, where measurements of outputs may be made from a variety of sources [11, 12], are of particular interest in the application of AL-ML to media design in order to fuse and correlate chemical assays that measure a similar target (proliferation of cells for example) but use different physical mechanisms (nuclear fluorescent labeling vs membrane infiltration for example). In later work [13] utilized a Bayesian data-fusion process, whereby a Gaussian process model correlated multiple long-term and short-term metrics of cell yield. Using such a data-fusion process, they were able to optimize a 14 component media system where the optimal medium generalized well to long-term cell growth (see figure below), which was not the case when using a single-source metric of proliferation in their previous work.

Multi-task modeling approaches [14], in which outputs are correlated with far less structure than multi-information models, may be used to bring together "omics" data in an AL framework to literately learn models to optimize metabolite concentration in culture media. Metabolomic or genomic data, such as nutrient consumption rates or gene expression, could improve the generalizability of ML models used in AL and could potentially be used to make inferences about causality between different tasks and metabolite concentration. This may require either more structured models informed by the biological interactions between omics layers, latent factor space representations so different datasets with different distribution or units may be fused, or perhaps deeper models such as neural networks or deep Bayesian kernels to learn more complex multi-omics relationships.

ML and the useful sub-field of Bayesian Optimization (BO) also includes tools such as active "causal" learning [15] in which experiments are designed such that causal networks are learned efficiently. For media design it may be valuable

to not just optimize a given economic / growth criteria, but to focus on learning significant interactions and effects for nutrients and growth factors. BO-based Information theoretic approaches [16] are also popular in finding optimal designs due to their focus on improving model and / or optimal point certainty. Multi-objective approaches [17], where multiple criteria are optimized at the same time, are also popular in the AL community (particularly in aerospace and engineering design) but much like information theory approaches, have not been applied extensively to biological design problems in general and media design in particular.

There seems to be an AL or BO method available for every conceivable design problem, with the primary obstacle, and greatest opportunity of value, being real-life use in laboratory or manufacturing settings. Therefore, it is critical that AL methods be developed with (or perhaps by) researchers working in wet labs. For example, it is unclear which architecture of models (GP, Student-t, deep models) are best suited to learning the relationships between different media components, nor how this changes with multi-omics data. There is no canonical way to incorporate experimental noise into ML models for biological systems so that AL methods can avoid overfitting while not ignoring valuable regions of the design space. Nor are there media design and optimization "test functions" that practitioners can use as baselines for their work to choose sampling methods, architectures, feature selection / regularization techniques, or optimizers. All of these questions must be studied but without real problem sets and in-depth knowledge of the biological / industrial domains to be optimized, investment in AL research is a mere academic exercise.

3 Active Learning and Gene Network Discovery

AL methods have applications in the discovery of gene regulatory networks (see figure below) that might be helpful in understanding the metabolic pathways for cell growth and differentiation. In [18] the authors used a random forest model to infer the metabolic pathway of aromatic amino acid pathway in yeast. The authors note that these methods could be extended to biologically similar organisms, unknown genes, and unknown enzymatic effects, which are valuable pieces of information in the development and scale-up of CA. Such applications could focus on the regulation of muscle cell proliferation via the IGF1-Akt/PKB pathway [19], where much work has been done to understand human skeletal muscle cells for health reasons but which could be repurposed to understand the metabolic needs and constraints of cells for CA industrial production. Similar work has also been done on identifying 'orphan enzymes', or enzymes which do not have an identified corresponding gene [20]. Much like the previous reference to "causal" AL methods for uncovering causality in networks, Bayesian continuous AL has been used to learn networks on test problem sets from biology [21] with applications in genomics. [22] has a good review of learning Boolean networks in the context of biological systems, with a case study in human cell proliferation signaling.

Unlike the application of AL to media design, network discovery of the interactions of the genome, metabolome, proteins, and the cellular environment have not attracted industrial use. However, many of the roadblocks related to cellular agriculture involve the need to maintain proliferation over long periods of time, differentiation potential into preferred tissues, robustness in bioreactor environments, genetic stability, and survival in non-traditional and suspension culture media. All of these research projects involve interactions of the cells with their various regulatory networks where AL methods could offer guidance in learning models, optimizing industrial conditions, and understanding the underlying processes in cellular agriculture. For more information, `https://sites.google.com/view/automatedscience/home/lectures/spring-2021#h.krn8jkpdckuz` has many applications and linked papers for AL and ML methods in biology including drug discovery, protein design, and chemistry.

4 Multi-omics and Proliferation / Differentiation

Multi-omics datasets, or the systems biology approach to integrating information about the genome, metabolome, proteome, among other bio-markers and bio-marker layers [23], also offers an area where ML can be used to further cellular agriculture. Here we take inspiration from the field of cancer biology where molecular features can be used to classify cancer phenotypes and sub-phenotypes. To take a single example of ongoing research, [24] discuss the difficultly of integrating data-types with different units, distributions, and co-variants to do classification of cancer types and sub-types. They address these issues with a variety of schemes from simple clustering using BIC regulated k-means and latent factor space representation to more complex but biologically derived structured models which would require intense collaboration between domain knowledge experts in biological pathways and computer scientists. For cancer in particular, low rank models to approximate the interactions between different sets of omics data have been used [25] but

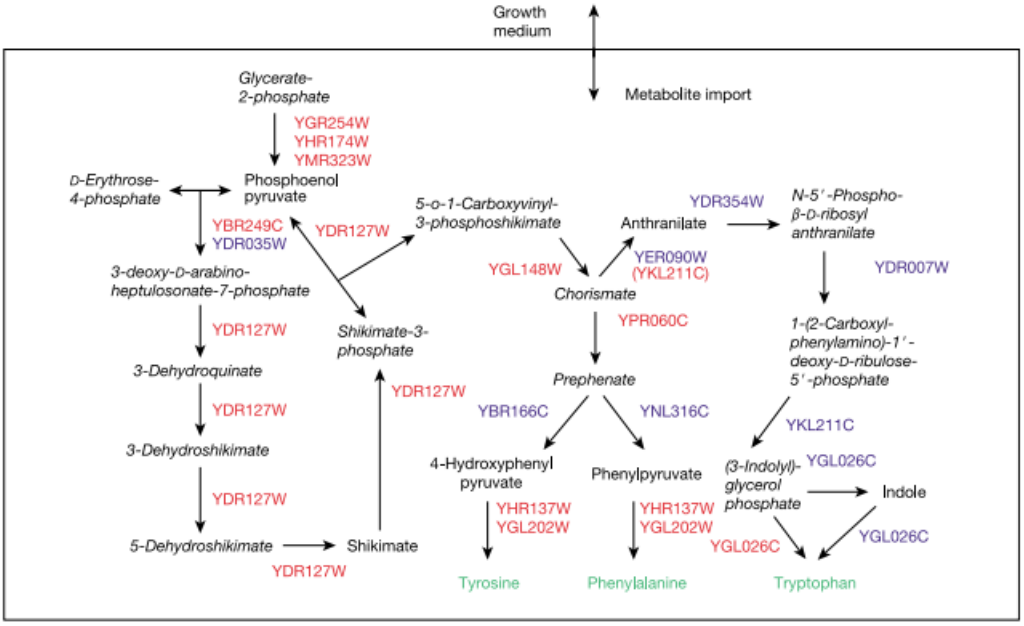


Figure 2: example of gene regulatory network for aromatic amino acid pathway in yeast provided by [18]

it is unclear is (i) how latent features can be connected to biological data [24] and (ii) whether the omics data and subsequent analysis can explain variation in the data (often it cannot) [26].

Such a paradigm could be applied to the analysis and optimization of cellular growth and differentiation, where ML researchers could help derive low dimensional feature representations of high dimensional and heterogeneous data. By measuring more biologically relevant features such as the transcription factors Pax7 or surface markers CD34, CXCR4, syndecan-4 and $\alpha 7$ integrin, ML may be used to better understand the complex causation of cellular proliferation (again, see [24]). Such methods could both complement and go beyond the AL methods discussed in the previous sections by (i) focusing AL approach in certain design spaces identified by multi-omics analysis and (ii) better understanding the causal mechanisms of media and bioreactor configurations. The same methodology could be imagined for phenotype analysis of cell types or multi-omics analysis of differentiation markers (for example see bio-markers at `https://www.rndsystems.com/research-area/myogenesis-markers` for muscle cells). The same challenges experienced in cancer biology related to accurate representation of biologically relevant latent factors and lack of data apply in CA as well. This militates the need for study in using ML to fuse multi-ommics datasets specifically for the outcomes, bio-markers, and phenotypes that are relevant for CA in particular.

5 Image Segmentation of Cells

Microscopy is one of the richest sources of data in cell culture, and thus finds value in CA, but it has historically been constrained by the complexity of its analysis. Microscopic analysis can assess important features of a cell culture, such as (i) the health of cells - e.g., whether they are mitotic, senescent, or apoptotic, (ii) the behavior of cells - e.g., whether they are invasive, contractile, or secretory, and (iii) the lineage of cells - e.g., whether they are stem cells, stromal or epithelial cells, terminally differentiated, etc. All of this analysis is routinely done manually by researchers with well-trained eyes, but doing it automatically requires systems that can incorporate many nuanced features of the image data. Unfortunately, appropriate systems tend to be nascent, poor, or non-existent in biological research. Fortunately, the challenges - image segmentation and classification - are well-defined, and there is a clear path to making them work better.

One of the most fundamental challenges in microscopic analysis is image segmentation, in particular dividing dense images of many cells into their constituent cells. This sort of segmentation is a prerequisite for microscopic analysis at

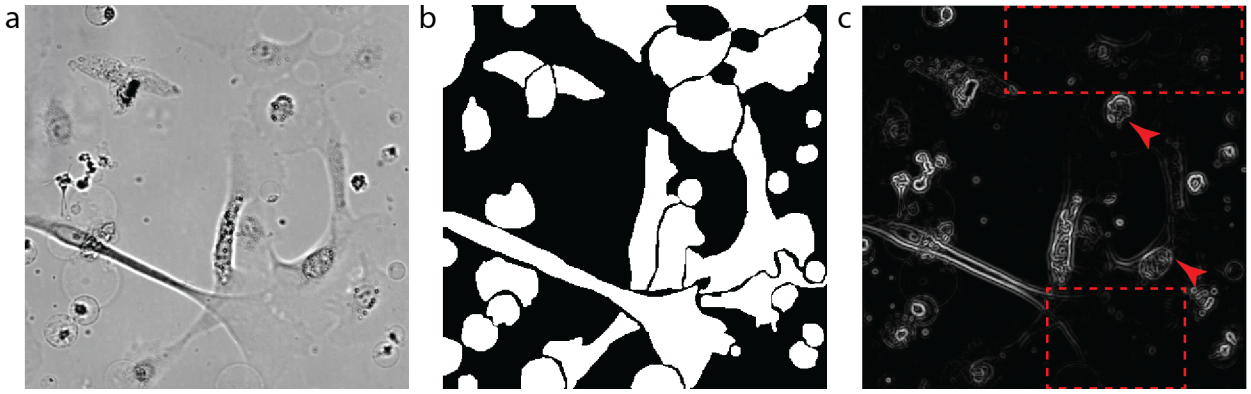


Figure 3: Traditional image segmentation has poor accuracy in brightfield microscopy. (a) Brightfield imaging of human mammary epithelial cells. (b) Manual segmentation. (c) Edge detection-based segmentation. Missing cells enclosed in dashed lines. Spurious segmentations indicated with arrowheads.

the single-cell level. When dealing with fluorescent or stained cells, segmentation is relatively easy because cytoplasmic or nuclear fluorescent dyes can brightly stain every cell, allowing segmentation by the watershed algorithm [27]. But when dealing with unstained cells, segmentation becomes a much harder edge detection problem wherein an edge of uneven brightness and thickness must be detected between a transparent background and a nearly transparent object. There are no widely accepted and reliable algorithms for unstained cell segmentation, and this is a bottleneck for microscopic analysis of cell growth, behavior, and morphology and could be an area of future research in the application of ML to CA when researchers and companies wish to be able to count and classify cells using fast, scaleable, and non-destructive techniques to better understand cell growth and morphology.

We emphasize unstained cell microscopy because of several practical benefits. Unstained cell microscopy is non-destructive (as opposed to histological staining or electron microscopy), which allows long-term, longitudinal imaging. Unstained cell microscopy does not require genetic or chemical manipulation, which simplifies workflows and, moreover, means that the toxicity of dyes and the side-effects of genetic transformation do not need to be taken into account. The most promising opportunity in microscopy image segmentation is probably training task-specific models based on convolutional neural networks (CNNs). CNNs are widely used for image segmentation and object detection in fields such as waste management [28] and self-driving cars [29], but the deployment of CNNs in microscopy is limited. At present, the majority of publicly available deep learning methods for microscopy are bundled in DeepImageJ [30], and, among these, U-Net is the most widely used architecture for microscopic image segmentation. U-Net, which is a biomedical CNN, was developed in 2015 [31] and had a readily deployable package published in 2019 [32]. However, U-Net is just an architecture, and it requires appropriately trained models in order to perform segmentation. Very few appropriately trained models are publicly available, even in repositories such as `BioImage.io` (`https://bioimage.io/#/`). Perhaps the single greatest point of leverage for improving microscopic image segmentation is training more models for CNNs such as U-Net. This could be done in collaboration with groups working on CA-relevant cell lines such as muscle and adipose cells.

6 Image Classification and Phenotype Analysis

Another fundamental challenge in microscopic analysis is classifying segmented cells by their morphological features to yield data. In brightfield microscopy, crude measurements such as cross-sectional area, circularity, and pixel intensity are the most commonly used features for this purpose. However, these measurements are rarely enough to classify important properties such as health, behavior, or lineage.

The challenge of microscopic image classification and the value of ML can be understood by looking to the field of pathology. Pathologists are medical specialists who diagnose disease largely by analyzing and classifying cell and tissue features from stained microscopic images. Notoriously, pathologists use subjective or nebulous criteria for image analysis, typically learned through on-the-job tacit knowledge and received wisdom from other pathologists. Although rubric-based criteria are sometimes used (and often desired) for pathological classification, they are typically inferior to subjective evaluation by a pathologist with "good eyes," owing to the subtle and varied microscopic features that

underpin pathology. Despite this difficulty, ML models can classify an increasing variety of microscopic images with accuracy comparable to fully trained pathologists [33, 34]. The field of pathology demonstrates that ML can overcome the difficulty in extracting classification data from microscopic images, potentially valuable as both a tool for researchers trying to conduct large-scale experiments of cell morphology and phenotype, and industrial end-users trying to monitor and control existing CA processes.

The most promising opportunity in microscopy image classification is probably, as with image segmentation, training task-specific models. Given appropriate segmentation, microscopic classification can be done with lightweight decision-tree models, such as random forests [35], as opposed to deep models such as CNNs. Random forests, despite their simplicity, have certain advantages over neural networks. (i) Random forests require less training data, which is important when building training data for rare lineages such as stem cells or rare events such as mitosis. (ii) Random forests are resistant to overfitting, which is important on segmented image data that might have dimensions or illumination different from the training data. (iii) Random forests have already been shown to work for microscopic classification [36]. The major barrier to employing random forests for microscopic classification is just producing supervised training datasets by people with appropriately trained eyes. With appropriate collaborate between ML and CA practitioners in generating large and well-curated datasets, image classification could become yet another tool in the arsenal of the growing CA industry.

7 Conclusion

In this short article we discussed some areas where the field of ML can aid in the development of CA. The disciplines of knowledge discovery (of media and regulatory networks) and of laboratory automation (using image classification or AL or design-of-experiments methods) have particular promise in this end. As a final note, the paucity of case studies of ML in the field should not be mistaken as indicative of a lack of promising applications, rather, the novelty of the field and the possibilities of greatly improving world food supply, animal ethics, and sustainability.

References

[1] C. Eldelman. Production of Cultured Meat. 11(5), 2005.

[2] Mark J. Post. Cultured meat from stem cells: Challenges and prospects. *Meat Science*, 92(3):297–301, 2012.

[3] David Humbird. Scale-up economics for cultured meat. Technical report, 2021.

[4] Edward N. O'Neill, Zachary A. Cosenza, Keith Baar, and David E. Block. Considerations for the development of cost-effective cell culture media for cultivated meat production. *Comprehensive Reviews in Food Science and Food Safety*, 20(1):686–709, 2021.

[5] Tatsuma Yao and Yuta Asayama. Animal-cell culture media: History, characteristics, and current issues. *Reproductive Medicine and Biology*, 16(2):99–117, 2017.

[6] Burr Settles. Active Learning Literature Survey. 2010.

[7] Zachary Cosenza, David E Block, and Keith Baar. Optimization of muscle cell culture media using nonlinear design of experiments. *Biotechnology Journal*, 16(11):2100228, nov 2021.

[8] Guiying Zhang, Matthew M. Olsen, and David E. Block. New experimental design method for highly nonlinear and dimensional processes. *AIChE Journal*, 53(8):2013–2025, aug 2007.

[9] Guiying Zhang and David E. Block. Using highly efficient nonlinear experimental design methods for optimization of Lactococcus lactis fermentation in chemically defined media. *Biotechnology Progress*, 25(6):NA–NA, 2009.

[10] Jan Havel, Hannes Link, Michael Hofinger, Ezequiel Franco-Lara, and Dirk Weuster-Botz. Comparison of genetic algorithms for experimental multi-objective optimization on the example of medium design for cyanobacteria. *Biotechnology Journal*, 1(5):549–555, 2006.

[11] Matthias Poloczek, Jialei Wang, and Peter I. Frazier. Multi-information source optimization. In *Advances in Neural Information Processing Systems*, 2017.

[12] Matthias Poloczek, Jialei Wang, and Peter I. Frazier. Multi-information source optimization, Supplementary Material. *Advances in Neural Information Processing Systems*, 2017.

[13] Zachary Cosenza, David E Block, Peter I Frazier, and Keith Baar. Multi - information source Bayesian optimization of culture media for cellular agriculture. (April):1–12, 2022.

[14] Kevin Swersky and Ryan P Adams. Multi-Task Bayesian Optimization. pages 1–9.

[15] Christian Toth, Lars Lorch, Eth Zürich, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius Von Kügelgen. Active Bayesian Causal Inference.

[16] Xiaokang Wang, Violeta Zorraquino, Minseung Kim, Athanasios Tsoukalas, and Ilias Tagkopoulos. Predicting the evolution of Escherichia coli by a data-driven approach. *Nature Communications*, 9(1), 2018.

[17] Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Multi-Fidelity Multi-Objective Bayesian Optimization: An Output Space Entropy Search Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):10035–10043, 2020.

[18] Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G.K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.

[19] Stefano Schiaffino and Cristina Mammucari. Regulation of skeletal muscle growth by the IGF1-Akt/PKB pathway: Insights from genetic models. *Skeletal Muscle*, 1(1):1–14, 2011.

[20] Andrew Sparkes, Wayne Aubrey, Emma Byrne, Amanda Clare, Muhammed N. Khan, Maria Liakata, Magdalena Markham, Jem Rowland, Larisa N. Soldatova, Kenneth E. Whelan, Michael Young, and Ross D. King. Towards Robot Scientists for autonomous scientific discovery. *Automated Experimentation*, 2(1):1–11, 2010.

[21] Hyunghoon Cho, Bonnie Berger, and Jian Peng. Reconstructing causal biological networks through active learning. *PLoS ONE*, 11(3):1–15, 2016.

[22] Nir Atias, Michal Gershenzon, Katia Labazin, and Roded Sharan. Experimental design schemes for learning Boolean network models. *Bioinformatics*, 30(17):445–452, 2014.

[23] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):1–15, 2017.

[24] Dongfang Wang and Jin Gu. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, 4(1):58–67, 2016.

[25] Marron JS Nobel AB Lock EF, Hoadley KA.

[26] Storey JD Chung NC. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 2015.

[27] Emad A. Mohammed; Mostaja M. A. Mohamed; Christopher Naugler; Behrouz H. Far. 2013.

[28] Chien-Tsung Wang Yu-Hao Lin Wei-Lung Mao, Wei-Chun Chen. 2021.

[29] Ling Guan Ahmed-Shaharyar Khwaja Abhishek Gupta, Alagan Anpalagan. 2021.

[30] Estibaliz Gómez-de Mariscal, Carlos García-López-de Haro, Wei Ouyang, Laurène Donati, Emma Lundberg, Michael Unser, Arrate Muñoz-Barrutia, and Daniel Sage. DeepImageJ: A user-friendly environment to run deep learning models in ImageJ. *Nature Methods*, 18(10):1192–1195, 2021.

[31] Wei Ouyang Laurène Donati Emma Lundberg Michael Unser Arrate Muñoz-Barrutia Daniel Sage Estibaliz Gómez-de Mariscal, Carlos García-López-de-Haro. 2021.

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

[33] Michael M. Franklin, Fred A. Schultz, Marissa A. Tafoya, Audra A. Kerwin, Cory J. Broehm, Edgar G. Fischer, Rama R. Gullapalli, Douglas P. Clark, Joshua A. Hanson, and David R. Martin. A Deep Learning Convolutional Neural Network Can Differentiate Between Helicobacter Pylori Gastritis and Autoimmune Gastritis With Results Comparable to Gastrointestinal Pathologists. *Archives of Pathology and Laboratory Medicine*, 146(1):117–122, 2022.

[34] A. Hekler, Jochen S. Utikal, Alexander H. Enk, Wiebke Solass, Max Schmitt, Joachim Klode, Dirk Schadendorf, Wiebke Sondermann, C. Franklin, F. Bestvater, Michael J. Flaig, Dieter Krahl, Christof von Kalle, Stefan Fröhling, and Titus J. Brinker. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, 118:91–96, 2019.

[35] LEO BREIMAN. Random Forests. *Machine Learning*, 45:5–32, 2001.

[36] Kun Hsing Yu, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher Ré, Daniel L. Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7:1–10, 2016.