Machine Learning in Cellular Agriculture: A Small-Data Approach
Akhil Jalan
January 2023

In 2023, the revolutionary impact of machine learning (also known as 'artificial intelligence,' 'data science,' etc) is nothing new. In the life sciences, AI has driven breakthroughs in protein structure prediction, medical image diagnosis, and drug discovery. Much of the progress in these areas is due to a "big data" approach where complex, multi-billion-parameter models learn complicated patterns from vast, human-labeled datasets. But these models are far too large for the working scientist to handle, and require teams of specialized data engineers just to maintain and store. Moreover, such models need humans to label billions of demonstrative data points by hand for the machine to learn patterns from. However, in emerging fields such as cellular agriculture, it is not yet clear what bioprocesses can be tractably modeled, let alone what data we may use to train our models.

Since cellular agriculture is currently a "small-data" problem, it needs a small-data approach to machine learning. In this article, I will explain one aspect of small-data machine learning, called *continuous learning.* Continuous learning refers to a knowledge-building process wherein researchers collect data from physical experiments, use those experiments to refine a statistical model, and then use the statistical model itself to guide further experimentation. This harmonious approach to physical and computational experimentation already has impressive results within cellular agriculture.

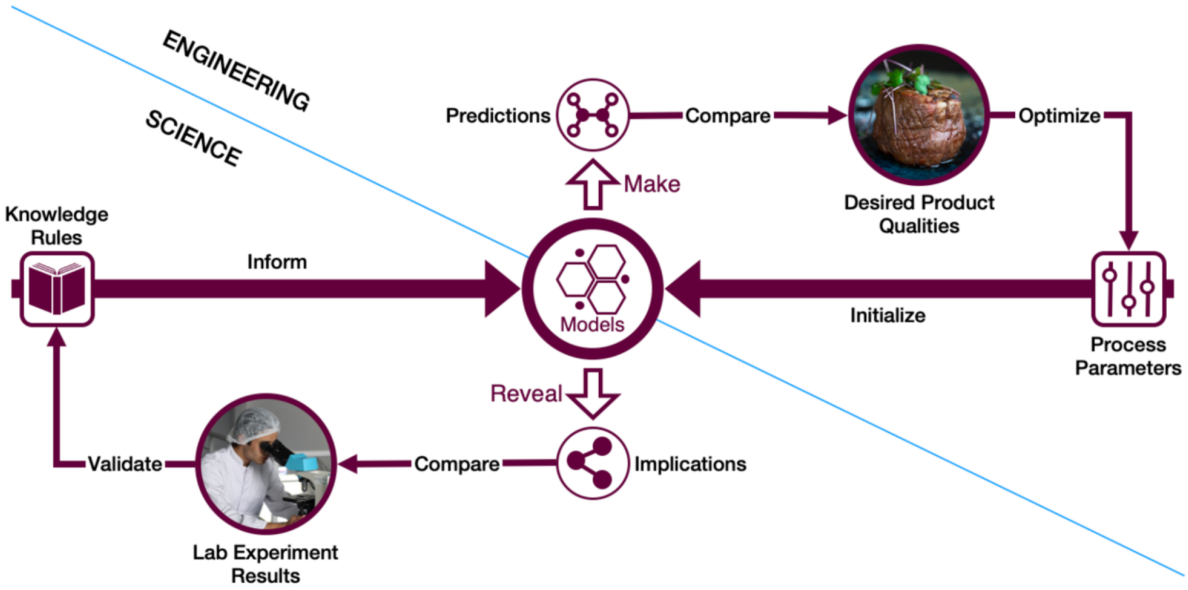

Figure: Models can inform science and engineering in a continuous feedback cycle. From the CMMC white paper.

For example, cell culture requires cost-effective media design for scale-up and commercialization. Cosenza et al. 2022 use a data-driven technique called Bayesian optimization (BO) to determine optimal concentrations of more than a dozen components of a cell culture medium for C2C12 cells. Their BO technique achieves 181% greater growth of C2C12 cells, and with 38% fewer experiments, than a baseline method. The key reason for their success is that they use a combination of cheap measurements of cell biomass (rapid chemical assays which can be done at scale) with more expensive, higher-quality measurements, in a principled statistical manner. This lets them achieve greater precision and results with much less data.

Small-data techniques are also effective for scheduling and automation problems in bioprocess development. For example, Bournazou et al 2016 optimize batch-fed growth of E. coli using robotic feeders and a continuous online optimization process that re-designs experimental conditions based on real-time observations. Using the same amount of physical resources, their techniques achieve 50 times greater precision (measured by the coefficient of variation) than the baseline sequential experimentation method. Sommeregger et al also highlight the importance of real-time monitoring and control for quality control in cell culture bioprocesses.

Future work should continue to use techniques from statistics and computational modeling to optimize bioprocesses in the presence of limited data. Collaborations between specialists in cell culture, biomanufacturing, and computational modeling are critical for such efforts to succeed. Large-scale consortia like the CMMC are therefore critical for the next generation of cellular agriculture researchers and practitioners to achieve their mission of a global food systems transformation.