Developing and Applying AI for Biopharmacology and Biomanufacturing

by

Pedro Seber

Submitted to the Department of Chemical Engineering
on April 23, 2026 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING

ABSTRACT

Biopharmaceuticals are an important class of drugs, highlighted by their high sales, recognition in awards, and the fact that biopharmaceuticals are the only or main treatment for certain diseases. Their prominence has been growing in recent years.

There are many open problems, both theoretical and practical, associated with biopharmaceuticals. One of the most important theoretical problems involves the occurrence and function of post-translational modifications. Glycosylation, a type of post-translational modification, is the addition of a sugar or sugars to a protein. These glycans have important structural and functional roles, but most studies of their function in biopharmaceuticals have been qualitative and non-systematic. For some glycan types, the glycosylation locations are not known without direct experimental measurement. Despite these knowledge gaps, glycans are essential for biopharmaceuticals to function properly and without toxicity, and are considered a critical quality attribute for biopharmaceuticals by the US FDA and the European Medicines Agency (EMA).

On the practical side, the main problems involve the optimization of bioproduction. Most biopharmaceuticals are produced within biological cells. Chinese hamster ovary (CHO) cells are the main cell used, responsible for producing 84% of all biopharmaceuticals in 2018. CHO cells dominate bioproduction because of their many advantages relative to other cells. However, CHO cells are inherently unstable, leading to productivity losses in the long term or the production of incorrect products. Although many causes of instability have been identified, no detailed mechanisms are available, preventing the use of mechanistic models. Furthermore, no data-driven models were available before my work.

Finally, another concern involves the modeling algorithms available to solve these problems. While many powerful tools, both mechanistic and data-driven, are available and have been applied to solve many scientific and engineering problems, it is possible that these tools are not sufficient to solve these biopharmaceutical problems with the available data. Moreover, the use of interpretable or explainable AI has been gaining prominence in many fields, especially those that deal with sensitive data or have critical applications. Models need to be not only powerful and accurate but also trustworthy and transparent, and they must be able to assist humans reliably, even in real scenarios with multiple failure modes.

This thesis presents discoveries and solutions to these and other problems primarily through the use of machine-learning, deep-learning, and hybrid-learning models. The tasks solved in

this thesis involve predicting glycosylation sites, quantitatively predicting N-glycosylation distributions, predicting metabolite levels over time in a bioreactor, and predicting long-term CHO cell stability. In addition, this thesis develops novel tools and algorithms to tackle machine-learning problems, with a focus on interpretable and explainable AI. These tools are compared to widely used methods and shown to perform considerably better in many problems investigated throughout this work.

Machine-learning, deep-learning, and hybrid-learning models have been seldom used in biopharmaceutical problems, in part due to challenges in obtaining sufficient amounts of high-quality data and in properly understanding how to train and use these data-driven and hybrid models. To address this second point, we developed an automatic machine learning (AutoML) software that can automatically train models and report their predictions and optimal hyperparameters without requiring any code or modeling-specific knowledge from the user. This software allows even those without any programming or machine-learning experience to train powerful models for a variety of tasks, including tasks unrelated to biopharmaceuticals and bioproduction.

Thesis supervisor: Richard D. Braatz
Title: Edwin R. Gilliland Professor of Chemical Engineering