Process Integration & Synthesis using Chemical Engineering Standards (PISCES)

PI: Corinne D. Scown, UC Berkeley

Problem

The scientific research and technology development needed to grow a robust domestic biomanufacturing sector in the United States requires a digital backbone capable of coordinating and leveraging data generated across multiple research institutions and activities, ranging from enzyme engineering, metabolic modeling, host engineering, to chemical process design. These researchers generate many different types of datasets that can be challenging to manage and learn from beyond the individual project level. Taking data from large numbers of individual studies and generating artificial intelligence (AI)-ready datasets is one of the grand challenges facing the field. Addressing this challenge requires balancing ease of sharing and storing data with the need to make these datasets practical to use beyond single-study analyses, culminating in a system that retains both the data itself and the nuances required for meaningful learning and predictions. Finally, the solution must prove to be useful through early applications that generate novel insights. Applying a standard format to end-to-end process designs documented in technoeconomic analyses can offer a compelling early application.

Solution

The challenges with storing and leveraging process data for biomanufacturing has parallels to the challenges being tackled in the development of autonomous laboratories for material synthesis and discovery. The Alab operating system (Alab OS) uses a directed acyclic graph format to store individual steps (called tasks) and dependencies (e.g., intermediate products moving from one task to the next) (Fei et al. 2024). We propose to:

1. Adapt and build upon this approach by developing a Standardized Flow Sheet Format (SFF) in which nodes correspond to unit processes in a simulated biomanufacturing facility and edges correspond to flows between those unit processes.
2. Using the SFF and leveraging large language models (LLMs), build a database of published process designs, including all relevant parameters for cash flow analysis and the final technoeconomic analysis results (CAPEX, OPEX). Different workflows involving different hand-off points between data extracted by LLMs and conventional calculations (e.g., cash flow analysis) will be explored to maximize comparability and accuracy. We will iterate on the SFF as needed to ensure it can accommodate the full range of relevant biomanufacturing processes. Several LLMs will be compared to identify the best-performing model(s) for extracting flowsheet data with minimal human intervention and hallucinated data.

3. Develop uncertainty analysis methods capable of tracking the robustness of empirical data underpinning process performance assumption, including potential scoring methods for distinguishing theoretical assumptions from performance that has been demonstrated at scale in representative conditions.
4. Identify and demonstrate potentially promising graph learning approaches for automatically completing incomplete process designs. Assess gaps in approaches and what data is needed to improve the quality of predictions.

The approach described here represents an early case study and proof of concept that can be further extended to track bench- and larger-scale processes, as is already implemented in the Alab OS. Methods for extracting data from documents such as experimental protocols in protocols.io and from other datasets stored and managed within BioMADE can ultimately be developed after this proof-of-concept is completed.

Contribution to Competitiveness

Establishing a robust digital backbone will be key to maximally leveraging BioMADE's investments in physical infrastructure to build a domestic bioeconomy. By using data to develop more automated analysis and design tools, it will be possible to speed up the pace of innovation and establish lasting institutional knowledge that can be built upon to secure the U.S.'s position as the leader in bioprocess innovation. Furthermore, the foundation built through this project can be used to disseminate both open-source large datasets and software, as well as platforms that can be leveraged by private entities to build their own internal knowledgebase in which practical process knowledge and lessons learned can be managed and shared across the organization.

The UC Berkeley team brings deep expertise in technoeconomic analysis (Scown et al. 2021), the realities of applying process design and technoeconomic analysis across a range of commercialization stages (Poddar and Scown 2025), the application of machine learning to technoeconomic analysis (Huntington et al. 2023), and in the development of interactive design tools (https://lead.jbei.org/tea_lca_tool/ ; https://dspdesigner.lbl.gov/ ; biositing.jbei.org). Through this partnership with team in BioMADE, it will be possible to combine this technical expertise with the BioMADE team's experience in developing software for large-scale applications to maximize the project's positive impact.

References

Fei, Yuxing, Bernardus Rendy, Rishi Kumar, Olympia Dartsi, Hrushikesh P. Sahasrabuddhe, Matthew J. McDermott, Zheren Wang, et al. 2024. "AlabOS: A Python-Based Reconfigurable Workflow Management Framework for Autonomous Laboratories." *Digital Discovery* 3 (11): 2275–88.

Huntington, Tyler, Nawa Raj Baral, Minliang Yang, Eric Sundstrom, and Corinne D. Scown. 2023. “Machine Learning for Surrogate Process Models of Bioproduction Pathways.” *Bioresource Technology* 370 (February): 128528.

Poddar, Tuhin K., and Corinne D. Scown. 2025. “Technoeconomic Analysis for Near-Term Scale-up of Bioprocesses.” *Current Opinion in Biotechnology* 92 (103258): 103258.

Scown, Corinne D., Nawa Raj Baral, Minliang Yang, Nemi Vora, and Tyler Huntington. 2021. “Technoeconomic Analysis for Biofuels and Bioproducts.” *Current Opinion in Biotechnology* 67 (February): 58–64.