Bioresource Technology 442 (2026) 133646

Contents lists available at ScienceDirect

Bioresource Technology

journal homepage: www.elsevier.com/locate/biortech

Development of multimodal AI for photobiorefineries via knowledge syntheses, transfer learning, and techno-economic analysis

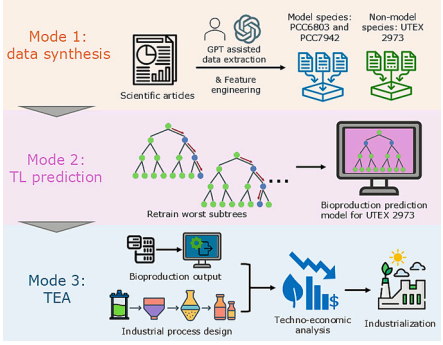
Runyu Zhao [a], Wenyu Li [a,b], Jose Gonzalez-Aguirre [c], Yayun Chen [d], Joshua S. Yuan [a], Himadri B. Pakrasi [e], Chengcheng Fei [d], Sunkyu Park [c], Garrett Roell [f], Yixin Chen [b], Yinjie J. Tang [a,*]

[a] Department of Energy, Environmental & Chemical Engineering, Washington University, 1 Brookings Drive, St. Louis, MO 63130, USA
[b] Department of Computer Science & Engineering, Washington University, 1 Brookings Drive, St. Louis, MO 63130, USA
[c] Department of Forest Biomaterials, North Carolina State University, 2820 Faucette Dr, Raleigh, NC 27607, USA
[d] Department of Agricultural Economics, Texas A&M University, 600 John Kimbrough Blvd, College Station, TX 77843, USA
[e] Department of Biology, Washington University, 1 Brookings Drive, St. Louis, MO 63130, USA
[f] Department of Molecular Biosciences and Bioengineering, University of Hawaii, 1955 East-West Rd., Honolulu, HI 96825, USA

HIGHLIGHTS

- Proposed first framework integrating LLMs, machine learning, and TEA for cyanobacteria.
- Constructed ML-ready datasets across UTEX 2973, PCC 6803, and PCC 7942.
- Enabled accurate prediction of UTEX 2973 bioproduction traits via transfer learning.
- Coupled AI models with TEA to reveal cost drivers in production.
- Accelerated DBTL cycles by streamlining biological performance to process economics.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
DBTL
GPT-4
Synechococcus elongatus
Photobiorefineries
Transfer learning
techno-economic analysis (TEA)

ABSTRACT

Synechococcus elongatus UTEX 2973 has emerged as a promising phototrophic chassis for next-generation biorefineries. However, the development of new cyanobacterial platforms is hindered by fragmented datasets, limited strain knowledge, and inadequate modeling tools. To address this challenge, we present MAP (Multimodal AI for Photobiorefinery), an integrated framework that combines large language models for knowledge synthesis, transfer learning for biological prediction, and techno-economic simulations for process evaluation. MAP employs GPT-4 and Graph-RAG (retrieval-augmented generation) to synthesize fragmented literature into coherent datasets and enable trustworthy knowledge retrieval. It further employs cross-species transfer learning from *Synechocystis* sp. PCC 6803 and *S. elongatus* PCC 7942 to improve the prediction power for growth and production traits of the emerging UTEX 2973 strain ($R^2 > 0.93$). Beyond data mining and performance prediction, MAP integrates model outputs with techno-economic analysis (TEA), enabling scenario evaluation across reactor types, cultivation conditions, and product pathways. Case studies in lycopene and sucrose production

* Corresponding author at: Department of Energy, Environmental & Chemical Engineering, Washington University, 1 Brookings Drive, St. Louis, MO 63130, USA.
E-mail address: yinjie.tang@wustl.edu (Y.J. Tang).

https://doi.org/10.1016/j.biortech.2025.133646
Received 7 October 2025; Received in revised form 3 November 2025; Accepted 11 November 2025
Available online 13 November 2025

highlight how MAP pinpoints cost drivers, informs strain engineering, and guides process optimization. Ultimately, this multimodal workflow bridges biological performance with process economics and accelerates the design-build-test-learn (DBTL) cycle for data-driven photobiorefineries.

1. Introduction

Artificial intelligence (AI) is shifting from a peripheral aid to a central driver of innovation in biotechnology and synthetic biology (Li et al., 2025). Across the DBTL cycle, AI accelerates nearly every stage: mining heterogeneous omics datasets (Flores et al., 2023; Xiao et al., 2023), predicting protein structures (Chen et al., 2024; Jumper et al., 2021), automating strain construction (Czajka et al., 2021; Radivojević et al., 2020), and guiding bioprocess optimization (Long et al., 2022; Mondal et al., 2023). Deep learning models increasingly capture latent design rules in biological systems and reduce experimental cycle times for both biomedical and industrial applications (Camacho et al., 2018). Consequently, AI is becoming embedded in standard workflows across bioengineering (Holzinger et al., 2023) and is regarded as a key enabler of next-generation biorefineries (Decker et al., 2023; Shah et al., 2025). However, despite impressive demonstrations in enzyme design, pathway engineering, and precision fermentation (Eslami et al., 2022), several barriers limit the full impact of AI-enhanced biomanufacturing. First, bioprocess data remain fragmented, sparse, and heterogeneous, limiting model generalizability and reproducibility across systems (Mateu-Sanz et al., 2024; Pei et al., 2022). Second, AI predictions often lack interpretability and causal grounding, leading to uncertainty, especially in non-model organisms (Ali et al., 2023; Lobentanzer et al., 2024; Zhang et al., 2016; Zhang et al., 2025). Third, even when strain traits can be predicted, commercial feasibility is governed not only by biology but also by economics. These challenges highlight the need for integrated frameworks that synthesize fragmented knowledge, predict outcomes for new hosts, and directly link biological performance with cost-driven feasibility analyses.

Among photobiorefinery candidates, cyanobacteria are particularly attractive because they fix CO_2 using sunlight while producing valuable chemicals (Knoot et al., 2018; Santos-Merino et al., 2019; Tan et al., 2022). Strains such as *Synechocystis sp*. PCC 6803 and *Synechococcus elongatus* PCC 7942 have been engineered for sugars, organic acids, terpenoids, and bioplastics under ambient conditions (Agarwal et al., 2022; Novoveská et al., 2023). More recently, *S. elongatus* UTEX 2973 has gained attention due to its remarkably fast growth, high photosynthetic efficiency, and potential for artificial consortium (Yu et al., 2015; Li et al., 2024; Zhao et al., 2022). Several integrative AI frameworks have recently emerged in biological synthesis, including BioTransformer for metabolic pathway reasoning and SynBioHub for data-driven genetic design (Djoumbou-Feunang et al., 2019; McLaughlin et al., 2018). While application of AI/machine learning algorithms in microalgae species inspires new opportunities for algal bioactive products (Guo et al., 2025). However, UTEX 2973 remains underdeveloped because published data are limited and fragmented, and no predictive framework links its biological potential with techno-economic viability.

This study develops a domain specific AI for cyanobacteria, MAP, by integrating Large Language Model (LLM) knowledge mining, cross-species transfer learning (TL), and techno-economic analysis (TEA). MAP structures published cyanobacterial studies into machine-learning-ready datasets, apply TL to predict growth and production outcomes for UTEX 2973 and couple these predictions with TEA to evaluate process economics. Case studies in lycopene and sucrose production demonstrate how MAP identifies key cost drivers, guides strain engineering, and accelerates DBTL cycles. By bridging predictive biology and economic feasibility, MAP provides a scalable, AI-driven workflow to advance carbon-negative cyanobacterial biorefineries.

2. Methods

2.1. Data collection and feature selection

Published literature was manually collected and organized into data features to build a comprehensive dataset for our ML models. This dataset spans three major categories: cultivation conditions, genetic engineering strategies, and product information. Cultivation conditions include parameters such as growth temperature, light conditions, reactor type and volume, initial culture volume and cell density, medium compositions, and environmental stress. Genetic engineering strategies encompass details on whether the strain is wildtype 2973, the type and concentration of promoter inducer, mutated and knocked-out genes, the method of gene insertion (either genomic integration or plasmid transformation), the gene of product exporter, strategies of cofactor enhancement, and whether the synthetic pathway is optimized for high yield. Lastly, product information encapsulates the product name, its enthalpy of formation sourced from the NIST database, atomic composition of the product molecule, precursor from central carbon metabolism, enzymatic steps for product synthesis from the precursor, and heterologous enzymatic steps from the precursor. The collected information was represented in two formats: numerical (e.g., cultivation temperature in ^{o}C) and categorical (e.g., reactor type represented as shaking flask, cylindrical photobioreactor, microfluidic droplet). Published information is often embedded in various formats within a document, such as text context, tables, and graphical figures. Graphical information was estimated as accurately as possible using a free online app (https://plotdigitizer.com/) that extracts data from images in numerical format. Similarly, the corresponding conditions were compiled based on the authors' evaluation of the published studies. During the first round of data collection, as many published details associated with strain growth and biosynthetic production as possible were recorded. This initial dataset was subsequently revised by removing interdependent features and merging single-reported genetic engineering strategies. This was done to avoid overfitting and to ensure ML robustness. Detailed information about the definition and importance of each feature is listed in the Feature Table.

2.2. GPT-4 assisted knowledge synthesis

This study leverages the power of advanced language models, specifically GPT-4 using a ChatGPT Plus account, to mine and structure data from a multitude of scientific papers.

Paper selection: For *Synechococcus elongatus* UTEX 2973, all bioproduction papers published up to the end of 2024 were screened. For the model strains *Synechocystis* sp. PCC 6803 and *S. elongatus* PCC 7942, the first round of papers was collected from the review "Cyanobacteria: Promising biocatalysts for sustainable chemical production." A second round supplemented this set with additional studies reporting bioproducts also synthesized in UTEX 2973, to ensure coverage and comparability. In total, 133 papers were curated.

Extraction procedure: The process was as follows: Firstly, basic information was extracted from abstracts (searched from PubMed), specifically focusing on product and strain names. Total number of 133 abstracts was processed. For example, prompts included: "Find the production of interest in the following study within three words" (for product name identification) and "List the strains of interest in the following study" (for strain identification). Next, the results sections were parsed to extract product concentrations. Given GPT-4's token limit (8192), the results were separated into shorter paragraphs.

Average tokens per query is around 7200. The prompt used was: "Highlight all the [product] production titers along with experiment conditions in the following text:", followed by the respective text. The extracted production information was then paired with cultivation conditions and strained metadata from methods sections or strain tables. Finally, GPT-4 generated a markdown table using the prompt: "Generate a spreadsheet from text with all of this information: strain | OD | product | production titer | gene information | cultivation time | temperature | light intensity | reactor type | reactor volume | culture volume | media | bicarbonate concentration | carbon dioxide concentration | pH | initial OD | inducer | inducer concentration.".

Error correction and hallucination detection: The GPT-4 generated tables were manually checked for mistakes and integrated into the database. Each datapoint (e.g., product titer, cultivation condition) was cross-verified against the original manuscript text, tables, or figures. Unit harmonization was applied to ensure metadata consistency. Hallucinations were defined as cases where GPT-4 generated values or conditions absent in the source; such entries were corrected or removed. This step relied on systematic human verification of every datapoint, rather than random sampling.

Validation metric: For the extraction stage, validation was defined as correctness of GPT-4 outputs relative to the source publications. Accuracy was evaluated by: (a) comparing extracted datapoints against the published content, (b) ensuring metadata integrity during manual curation, and (c) benchmarking Graph-RAG retrieval.

Graph-RAG integration: LLMs are prone to hallucinations and lack grounding. Retrieval-Augmented Generation (RAG) strategies help mitigate these limitations. In this study, Graph-RAG software (https://microsoft.github.io/graphrag/) supported trustworthy knowledge synthesis, enabling citation-linked retrieval and internal query traceability. The graph-based retrieval layer, benchmark results, and example prompts are provided additionally.

2.3. Genome scale model simulated production

Constrained-based analysis was employed on COBRApy v0.20.035 in Python v3.13.5, and the packages Pandas v2.3.1, NumPy v2.2.6, Matplotlib v3.10.3. The metabolic models (iJN678 for PCC 6803, iJB785 for PCC 7942, and iSyu683 for UTEX 2973) were used for the computational analysis (Broddrick et al., 2016; Mueller et al., 2017; Nogales et al., 2012). For the simulation of recombinant cyanobacteria strains producing metabolites of interest, the metabolic reconstruction was expanded with unidirectional pseudo-reactions allowing product export of the respective compounds (available on GitHub page). For new reactions to the models, the reaction and metabolite information was obtained from BIGG and MetaCyc (Caspi et al., 2020; King et al., 2016). All pseudo-reactions were mass- and charge-balanced based on metabolite formulas from these databases, and stoichiometric consistency was confirmed prior to analysis.

To evaluate the maximum production rate of the metabolite of interest under specified growth rates, flux balance analysis (FBA) was employed (Orth et al., 2010). Flux values are reported in mmol $gDW^{-1}h^{-1}$, consistent with COBRApy conventions. Exchange fluxes were constrained using literature values where available: CO_2 uptake was fixed at 3.7 mmol $gDW^{-1}h^{-1}$ (Kugler et al., 2023), and photon uptake was set to 100 mmol $gDW^{-1}h^{-1}$ for both photosystem reactions. Other exchange reactions (e.g., O_2) were assigned wide bounds (−1000 to + 1000 mmol $gDW^{-1}h^{-1}$) to allow reversible transport unless otherwise specified. Pseudo-reactions for heterologous product secretion were defined between 0 and 1000 mmol $gDW^{-1}h^{-1}$ to permit export but not uptake. The minimum flux through the biomass objective function was set to 10 % of the predicted maximum growth rate.

2.4. Data curation and processing

Data was sourced from two primary Excel files. The first two files, 6803.xlsx and 7942.xlsx (the source dataset), and the third, 2973.xlsx (the target dataset). All files were uploaded in a directory labeled./data/ . Specific columns were extracted for analysis, excluding any column named 'Unnamed: 0'. Entries listed as 'na' or blank were treated as missing values. Missing data were imputed using a set of predefined rules outlined in an additional Excel file named impute.xlsx. This file contained rules specifying the imputation technique to be applied to each feature. Implemented imputation methods included filling missing values with the average (for numeric data), the mode (for categorical data), zero, or treating non-appliable entries as a separate category ('NA'/'na'). Out of all the data entries, 19.2 % is imputed using the imputation method. The imputation process was applied separately to both the source and target datasets. For model training and validation, specific features were designated as either input or output variables. Output features included 'OD', 'growth rate', 'product titer', and 'production rate'. All remaining features were considered as input features. Categorical variables within the datasets were encoded using an Ordinal Encoder to transform them into numerical format. This was followed by scaling all features to a range between 0 and 1 using a MinMaxScaler to normalize the data, ensuring uniformity in magnitude which is a prerequisite for many machine learning algorithms. The training and testing datasets were constructed based on a unique identifier, 'paper', presumed to delineate distinct experimental batches or groups within the data. A random selection of three identifiers was used to create the test dataset, with the remainder used for training. This methodological approach ensures a robust framework for handling, processing, and analyzing the data with minimal information leakage, adhering to standard practices in data-driven modeling to maintain the integrity and reproducibility of the research findings.

2.5. Machine learning model training and evaluation

Machine learning and transfer leaning was performed in Python v3.10 using the packages Scikit-learn v1.6.1. A Random Forest (RF) Regressor with TL is employed to predict the designated output features based on the input features. Grid search with 5-fold cross validation is first used to rigorously explore the parameters space by optimizing Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Mean Arctangent Absolute Percentage Error (MAAPE) by considers the percentage error:

$$\mathrm{MAAPE} = \frac{1}{n}\sum_{i=1}^{n}\tan^{-1}\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right)$$

and R-squared (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

which provide measures of the model's prediction accuracy thoroughly, independent of the scale or units of the features. The grid search ranges combinations of number of estimators (100, 200, 500) and max depth (5, 10, none). The TL approach involves refining the RF regressor by selectively retraining the weakest-performing decision trees using a combined dataset with source data. The numbers of weakest-performing trees retrained are 1, 3, 5, 10, 15, 20, 25, 30. Initially, individual tree performance is evaluated on the target test data, with R^2 scores used to rank the trees in ascending order. The method iteratively retrains a varying number of the worst-performing trees using a dataset that combines source and target data, thus leveraging source knowledge to improve target domain predictions. After retraining, the updated RF model is evaluated on the target test set, and the configuration yielding the highest R^2 score is identified. TL efficiently adapts the RF model to

the target domain by focusing on underperforming components of the ensemble and optimizing predictive performance.

2.6. Techno-economic analysis

TEA was performed following established guidelines from Perry's Chemical Engineering Handbook (9th edition) and methodologies applied in prior TEA studies for algal and photobioreactor systems (Banerjee & Ramaswamy, 2019; Davis et al., 2016; Zhu et al., 2018). The objective was to estimate the minimum selling price (MSP) of sucrose and lycopene produced using UTEX 2973 under various cultivation scenarios, with input parameters informed by machine learning predictions of product titer, biomass concentration, and growth rate.

Process Modeling and Input Parameters: The mass balance was derived from product titer predictions generated by TL and literature values, while energy balances (electricity, cooling, and process energy demands) were sourced from studies (Banerjee & Ramaswamy, 2019; Zhu et al., 2018). Process simulations were conducted for both batch-mode sucrose production and semi-continuous lycopene production across different CO_2 concentrations (0.5 % and 5 %), cultivation durations (120 and 240 h), and reactor volumes (0.1 to 1000 L). For certain downstream processes, Aspen Plus V12.1 was used for modeling at a base reactor volume of 1,250 m^3. The production cost basis was set for the year 2024 (some cost parameters were calculated from 2023 reference data).

Capital Cost Estimation: Capital costs were scaled using cost correlations from Perry's Handbook and cost models from the National Renewable Energy Laboratory (NREL) photobioreactor reports. For sucrose production, major equipment included cylindrical photobioreactors, CO_2 delivery and storage systems, dewatering units, and storage tanks. For lycopene production, major equipment comprised helical tubular photobioreactors (PBRs), inoculation systems, CO_2 storage and piping, makeup water systems, dewatering equipment, storage tanks, and lycopene extraction units. Lycopene extraction costs were informed by multiple solvent-assisted extraction studies, including those on tomato residues (Casa and Miccio, 2021), food- and microalgae-based lycopene recovery (Martins et al., 2023), and comparative evaluations of green extraction and purification processes (Yadav & Dhamole, 2024), with adjustments made to reflect the solvent recycling efficiency and physicochemical characteristics of cyanobacterial biomass. Additional capital components such as piping, fittings, pumping, and electrical infrastructure were referenced from Zhu et al. (2018). Indirect capital costs accounted for engineering supervision (5 %), installation (10 %), and insurance and taxes (1 %) of the total direct capital costs.

Operating Cost Estimation: Variable operating costs included CO_2, ammonia, diammonium phosphate (DAP), electricity, chiller utilities, salt disposal, lycopene extraction solvent (hexane with recycle), and makeup water. Fixed operating costs included labor (based on required employee salaries), maintenance (5 %), overhead (10 %), and administrative expenses (10 %), each expressed as a percentage of the total direct capital costs. Market prices for chemical inputs were obtained from online industrial suppliers like Alibaba (https://www.alibaba.com/) and ChemCentral Marketplace (https://www.chemcentral.com/), and adjusted using standard chemical price indices where necessary.

Economic Evaluation: The production cost was calculated as the sum of fixed capital investment and total annual operating costs. The MSP was determined as the unit product price required to achieve a net present value (NPV) of zero over the project lifetime using a discounted cash flow approach. Annual production of lycopene was assumed at 100 metric tons across all manufacturing conditions for comparative consistency. Sensitivity analysis was performed to evaluate the impacts of key parameters (e.g., CO_2 concentration, product yield, electricity price) on MSP, identifying major cost drivers and potential design optimizations.

3. Result and discussion

3.1. Dataset construction and feature organization

High-quality datasets are essential for training reliable AI/ML models in bioproduction. To build such a resource for cyanobacteria, we compiled information from 133 peer-reviewed publications, covering more than 2,100 engineered strains of Synechococcus elongatus UTEX 2973, Synechococcus elongatus PCC 7942, and Synechocystis sp. PCC 6803 by the end of 2024. The dataset spans a wide range of product classes, including alcohols, carbohydrates, organic acids, lipids, terpenoids, and bioplastics, together with cultivation parameters (e.g., temperature, light intensity, reactor type, nutrient composition), genetic modifications, and product performance metrics. Theoretical yields were also included by simulating engineered pathways with genome-scale models (iSyu683 for UTEX 2973, iJN678 for PCC 6803, iJB785 for PCC 7942) (Broddrick et al., 2016; Mueller et al., 2017; Nogales et al., 2012). These mechanistic benchmarks complement the experimental data, grounding the predictions in biological feasibility.

Overall, the dataset captures a broad performance spectrum, with optical density values ranging from 0.03 to 45 and product titers from 0 to 10 g/L. Four major outputs (OD, growth rate, product titer, and production rate) were selected for model training. To construct this resource, we applied a hybrid workflow: GPT-4 extracted relevant information from full-text publications into multi-column tables, while manual curation ensured unit consistency and accurate labeling. The finalized dataset is now publicly available on IMPACTDB (https://impact-database.com/), improving accessibility and supporting reproducibility across the field (Fig. 1). By unifying previously fragmented data into an ML-ready resource, this dataset not only enables predictive modeling but also allows microbiologists and engineers to identify common design strategies and bottlenecks in cyanobacterial engineering efforts.

3.2. Transfer learning (TL) from PCC7942 and PCC6803 to UTEX 2973

A key challenge for predictive modeling of UTEX 2973 is the limited amount of strain-specific data. To overcome this, we applied TL using knowledge from two better-studied cyanobacteria, PCC 7942 and PCC 6803. The goal was to improve predictions of growth and production traits in UTEX 2973 by leveraging similarities with related species. The TL models were trained to predict four main outputs: optical density (OD), growth rate, product titer, and production rate. Bar plots (Fig. 2A) summarize performance metrics before and after TL. Using random splits of the dataset (70 % training and 30 % training), baseline predictions already showed high accuracy ($R^2 > 0.92$), partly due to overlapping feature values between training and testing sets. Even in this scenario, TL consistently improved accuracy, particularly for product titer and production rate. For example, OD predictions improved from $R^2 = 0.93$ to 0.94, and product titer predictions from $R^2 = 0.95$ to 0.96.The strongest improvements came from transferring knowledge from PCC 7942, which is nearly identical to UTEX 2973 at the genomic level (99 % ANI, differing by only 55 SNPs, a 7.5-kb indel, and a 188-kb inversion) (Yu et al., 2015; Ungerer et al., 2018; Adomako et al., 2022). In contrast, PCC 6803 is more distantly related and knowledge transfer from this species provided smaller gains, likely due to differences in stress responses and metabolic regulation (Billis et al., 2014; Stork et al., 2005). Combining both PCC 6803 and PCC 7942 data did not outperform PCC 7942 alone, suggesting that excessive heterogeneity may dilute the predictive power.

Scatter plots (Fig. 2 B - E) further illustrate that TL improves model predictions to align better with experimental data. The models were more accurate for fast-growing and high-producing strains, while slow-growing and low-producing strains deviated more strongly. Because the low-titer strains exhibit more unmodeled bottlenecks (Czajka et al., 2021). For example, myo-inositol production in UTEX 2973 burdens

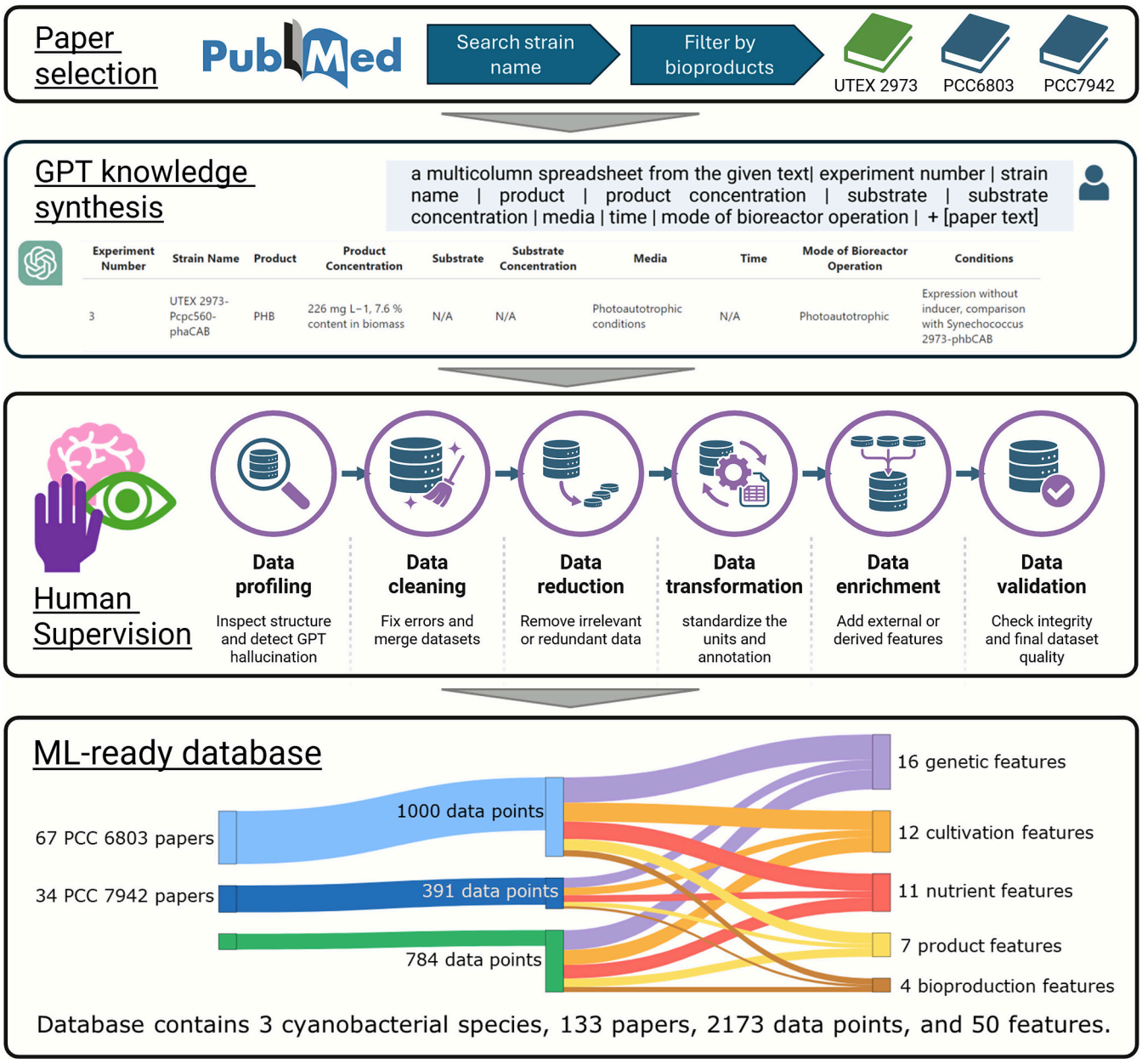


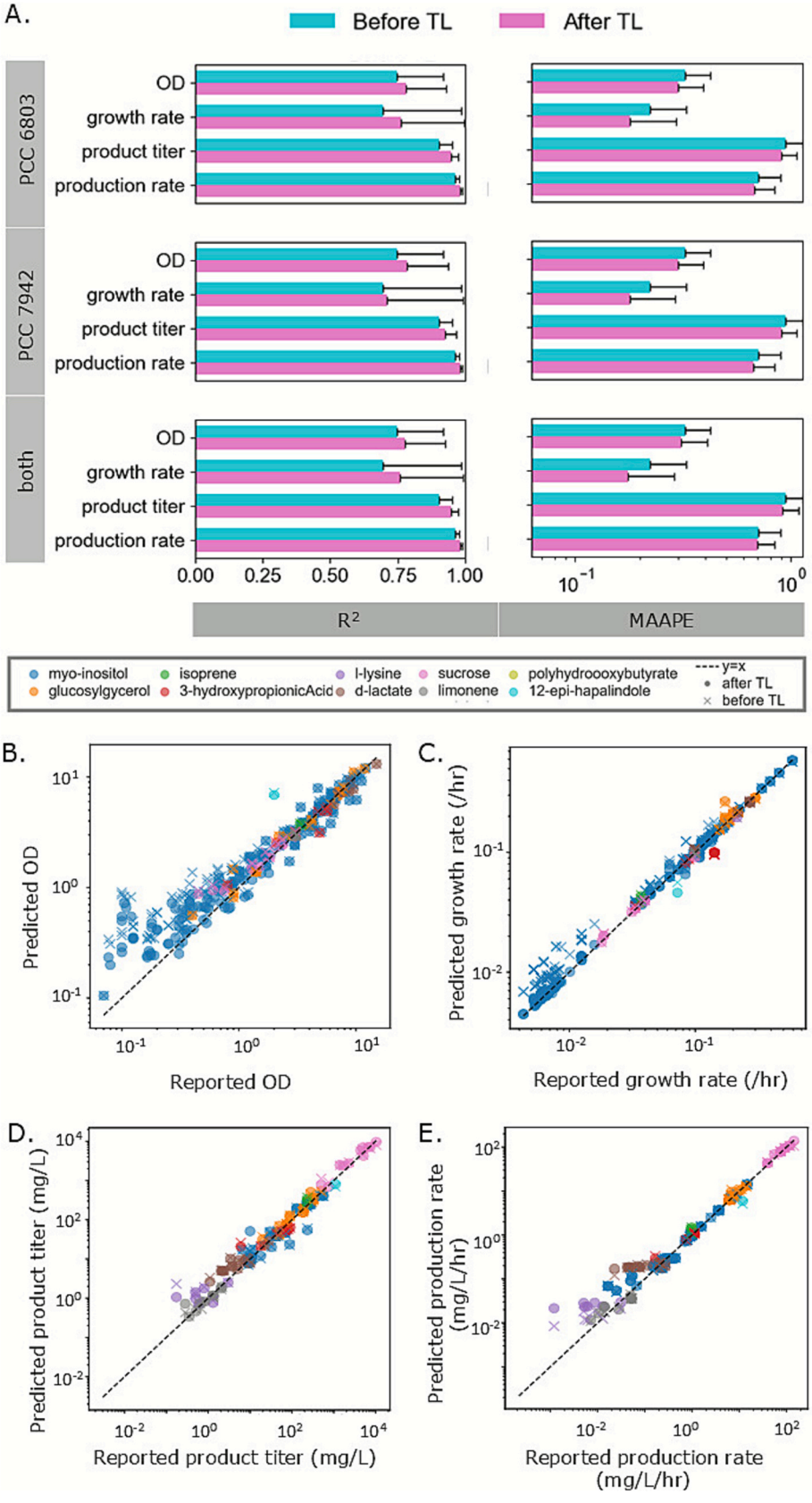
Fig. 1. Demonstration of the data mining process powered by gpt4 followed with human supervision and resulting to ML-ready database.

biomass growth by the inositol-3-phosphate synthase diverting glucose-6-phosphate from central metabolism and depleting NADH/NAD + pools, leading to growth-product trade-offs (Sun et al., 2023). Biased reducing power compromises cyanobacteria growth performance and adds complexity that lacks representation among the current features. These factors explain the challenges in modeling low-growth, low-OD strains and highlight areas for future feature refinement.

In summary, TL enabled accurate prediction of UTEX 2973 traits by leveraging data from related species, particularly PCC 7942. This demonstrates how computational models can bridge data gaps for emerging chassis strains and provide actionable guidance for metabolic engineering and process optimization.

3.3. Stringent TL test: Split data by individual studies

While the random-split results (Section 3.2) demonstrated strong model accuracy, these tests allowed data from the same study to appear in both training and testing data sets. This information overlap can artificially inflate performance because similar cultivation conditions or genetic modifications may recur within a single paper. To more closely mimic real-world applications of predicting new experimental setups, we performed a stringent test by splitting the ML-ready dataset according to individual publications. Three papers were randomly selected from the UTEX 2973 dataset as test data. Under this study-based split, baseline model performance dropped substantially, with R^2 values as low as 0.10 ~ 0.27 for OD and production rate and very high mean squared errors (MSE > 10^6) for product titer. This confirmed that models trained without TL struggled to generalize to unseen biological contexts for novel strains like UTEX 2973, for which limited biosynthesis study were performed and reported. Applying TL, however, dramatically improved predictive power. R^2 values increased to 0.71 ~ 0.82 for OD, 0.91 ~ 0.93 for product titer, and 0.78 ~ 0.83 for production rate, while MSEs were significantly reduced (Fig. 3). These gains highlight the importance of leveraging relevant prior knowledge when experimental data are scarce. Agreeing on the results of random-split in Section 3.2, transferring knowledge from PCC 7942 produced the best overall



(caption on next page)

Fig. 2. UTEX 2973 predictive model performance. Bar plots in (A) panel depict the R^2 and MAAPE before and after TL using source data of PCC 7942, PCC 6803, and both species. Error bars depict the 95 % confidence interval calculated from five independent random train/test splits. (B–E) Scatter plot illustrating the accuracy of model predictions compared to experimentally reported values for: (B) optical density (OD), (C) growth rate, (D) product titer, and (E) production rate. The R^2 scores before and after transfer learning were: OD (0.90 → 0.91), growth rate (0.81 → 0.81), product titer (0.81 → 0.94), and production rate (0.92 → 0.94), respectively. Different metabolites and compounds, including myo-inositol, isoprene, l-lysine, sucrose, polyhydroxybutyrate, glucosyl glycerol, 3-hydroxypropionic acid, d-lactate, limonene, and 12-*epi*-hapalindole, are distinguished by color. Data points represent conditions before (cross) and after (circle) TL. The dashed line indicates perfect prediction (y = x).

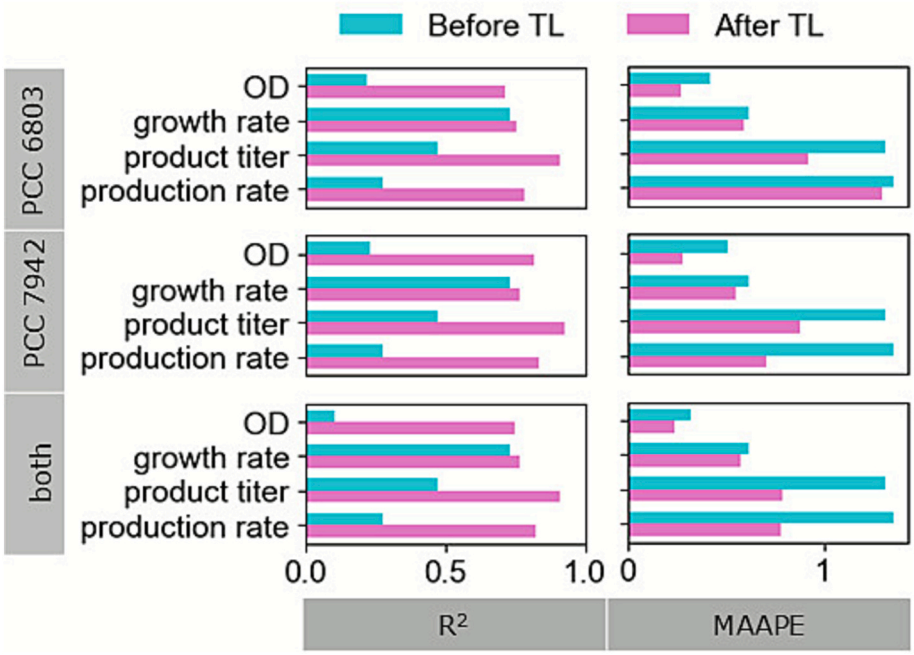


Fig. 3. Stringent test result of UTEX 2973 predictive model. The bar plots depict the R^2 and MAAPE before and after TL using source data of PCC 7942, PCC 6803, and both species.

performance, outperforming both the PCC 6803-only and the combined transfer approach (Fig. 3). The genetic similarity between the knowledge source strain and prediction target strain still plays an important role in stringent tests. Knowledge of PCC 7942, the genetic sister of UTEX 2973, is more valuable in TL than that of PCC 6803. This consistent observation in both random-split test and individual study-split test can inspire effective data collection for future biological ML studies. This stringent test demonstrates that TL can recover predictive accuracy even under realistic conditions where there is no overlap between the training and testing data. For microbiologists and bioprocess engineers, this result underscores that computational models informed by related species can provide reliable guidance for strain engineering and scale-up decisions, even when direct experimental data are limited.

3.4. *Feature importance analysis of TL prediction models*

To translate TL results into biology-related insights, feature importance analysis was performed for all the tested models. The analysis revealed that the predictive models relied heavily on a relatively small subset of features. More than 80 % of the features had importance values below 0.05 out of 1, suggesting that the majority of inputs contributed little to model performance. Instead, a smaller group of features with importance values above 0.1 explained most of the predictive power. These highly weighted variables included cultivation time, CO_2 concentration, initial OD, NaCl concentration, and inducer concentration, along with several genetic design parameters such as the number of heterologous enzymatic steps, biosynthetic precursor, knockout genes, and the presence of a product exporter. Product-related features also ranked among the most influential, particularly product identity, atomic composition (e.g., number of carbon and hydrogen atoms), and product enthalpy of formation (Fig. 4A). When comparing models before and after TL, most features retained similar levels of importance (Fig. 4A). Meanwhile, some variables showed notable shifts that explain the performance gains achieved through TL. Product-related information became more influential after TL, reflecting the importance of metabolic characteristics in determining both biomass accumulation and product synthesis. Cultivation time remained one of the strongest predictors, but its high weight also highlighted a limitation of the current dataset, which does not capture the full temporal dynamics of cyanobacterial metabolism. NaCl concentration gained prominence after TL, consistent with the sensitivity of cyanobacteria to osmotic stress and the strong impact of salinity on growth and productivity (e.g., sucrose). By contrast, factors such as stress type and product enthalpy of formation were relatively less important due to inconsistent reporting and limited variation across the compiled studies. Feature importance analysis also revealed that features contributed differently depending on the prediction task (Fig. 4B). For growth-related outputs (OD and growth rate), cultivation features such as cultivation time, CO_2 concentration, and initial OD dominated the predictions. For bioproduction outputs (product titer and production rate), product information and genetic engineering features played a more decisive role, emphasizing the central influence of metabolic design choices and product properties on production performance.

Feature importance also provided insights into the role of information overlap in shaping predictions. As illustrated in Fig. 4B, random data splits tended to mask the importance of some biologically relevant variables. For instance, the effect of CO_2 concentration on growth rate was underestimated when overlaps occurred, while for production predictions, the influence of stress type, knockout genes, initial OD, biosynthetic precursor, and product exporter was relatively neglected.

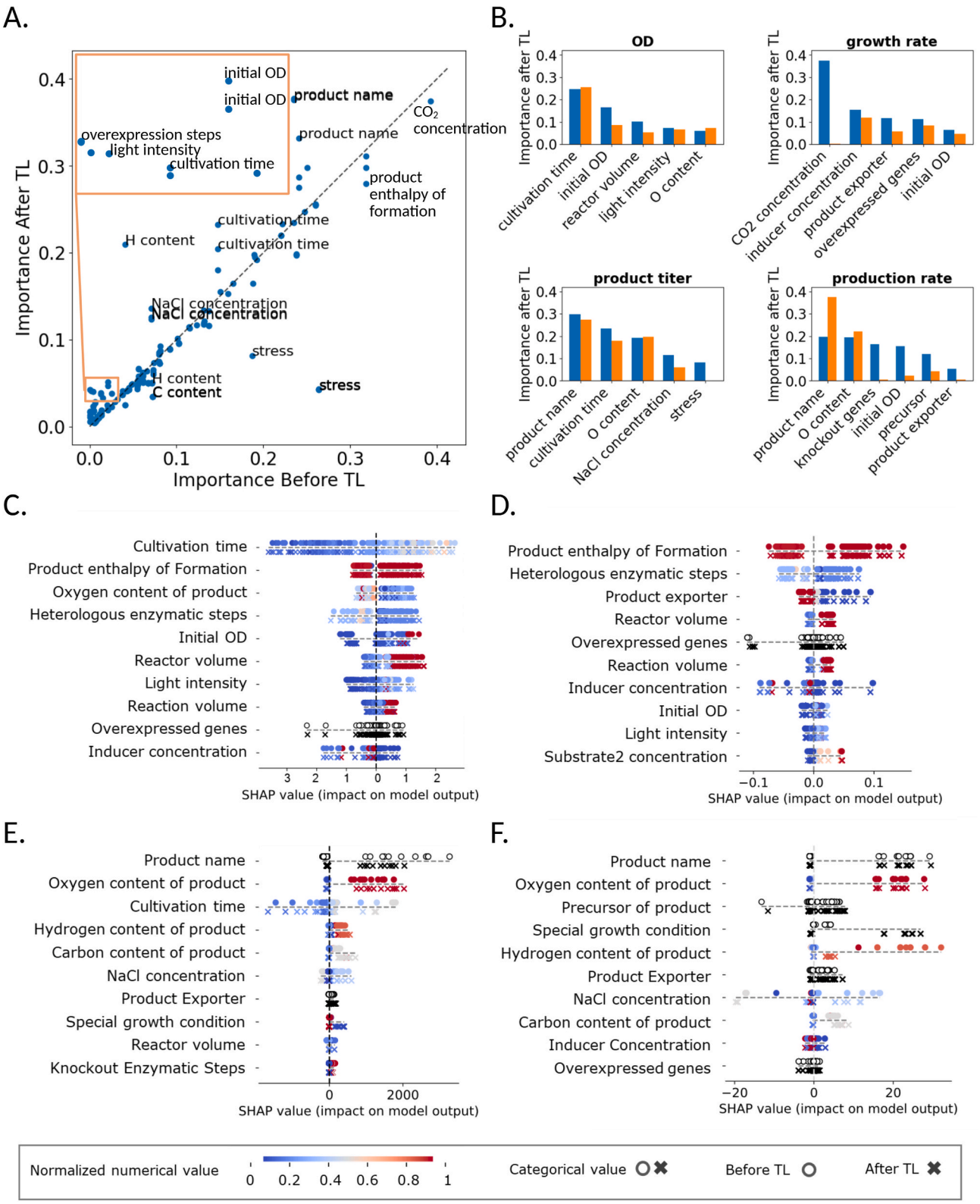


Fig. 4. Feature importance analysis results. (A) Feature importance shift before and after TL. The dash line indicates no importance shift of the feature after applying TL. (B) Feature importance after TL by different data splitting rules. Plots only included features with importance value over 0.1. (C-F) SHAP summary plots showing the top ten features of model predictions for four separate machine-learning tasks. Each panel displays the ten features with the highest mean absolute SHAP value for model predicting (C) OD, (D) growth rate, (E) product titer, (F) production rate. Along the y-axis features are ordered by importance from top to bottom; the x-axis gives the SHAP value (i.e. the effect on the model output, with positive values pushing predictions higher and negative values pushing them lower). Each point represents one observation's SHAP value, colored by the normalized numerical feature value (blue = low, red = high) and shaped (○ before TL or × after TL) by categorical level.

In contrast, under the more stringent study-based split, these features emerged as more relevant. This demonstrates that information overlaps in model training can distort the apparent weight of biological variables, leading to misleading conclusions.

The compilation of literature-derived data inevitably introduces biases, such as the underrepresentation of low-performing strains and inconsistent reporting of experimental details. For instance, certain stress types and product enthalpy of formation were less important in the feature analysis, which we attributed to inconsistent reporting and limited variation in the compiled studies. Conversely, features that were consistently well-reported, such as cultivation time and CO_2 concentration, naturally received higher weights in the model. The dominance of cultivation time as a strong predictor, while biologically relevant, also highlights a limitation of our current static dataset, which may not fully capture the metabolic dynamics, further suggesting a data-structure-related influence on the feature weights. To further validate the robustness of feature contributions, additional SHAP (SHapley Additive exPlanations) value plots confirmed that the same small set of features (cultivation time, CO_2 concentration, initial OD, NaCl concentration, and product atomic composition) were the most influential drivers for predictions of OD, growth rate, product titer, and production rate (Fig. 4C-F). This agreement between TL feature importance and SHAP values reinforces the conclusion that these variables are genuinely the most significant biological and process parameters captured by our current dataset for predicting UTEX 2973 performance.

Taken together, the feature importance analysis indicates that a few key features (e.g., product identity, cultivation conditions, and genetic modifications) drive most of the predictive accuracy. These findings not only explain the boost in performance after TL but also point to priorities for future dataset construction. By focusing on well-reported, high-impact features, future algae biosynthesis studies can reduce redundancy in feature space while improving both predictive modeling and biological interpretability.

3.5. TEA integrated with TL predictions

To assess the economic potential of UTEX 2973, an AI-integrated techno-economic analysis framework was integrated to combine TL predictions with cost modeling to evaluate the economic feasibility of bioproducts. The framework begins with predictive models trained via TL, which forecasts key bioproduction features including final biomass concentration, specific growth rate, product titer, and production rate, across diverse process conditions. These predictions are then converted into reactor-scale annual yields in batch and semi-continuous modes. The resulting output informs the TEA model, enabling estimation of unit production costs ($/kg) under different reactor configurations and cultivation strategies. The PBR size and mode determines part of the capital cost. The predicted OD can be converted to biomass yield and provide parameters for the CO_2 storage and piping. The nutrient features provide parameters for nitrogen and phosphorous consumption. Other parameters were adapted from literature or obtained through simulation in Aspen. Simulations were conducted using either helical tubular PBRs operated in a semi-continuous mode for lycopene production, spanning a matrix of CO_2 concentrations (0.5 % and 5 %), cultivation durations (4

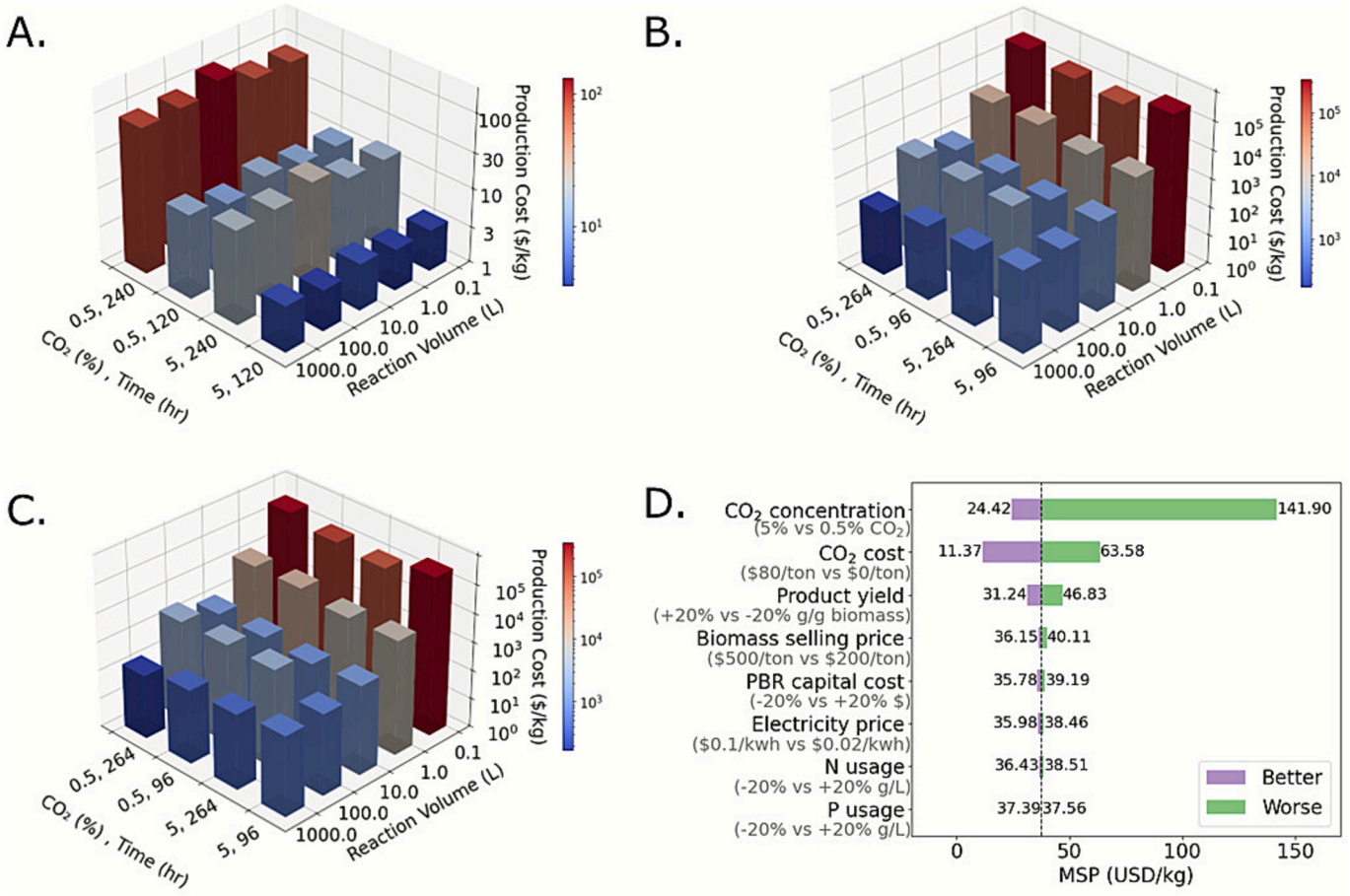


Fig. 5. Optimization and sensitivity analysis of cultivation parameters for economic production of lycopene. (A-C) 3-dimensional bar plots showing the production costs of lycopene using (A) semi-continuous helical tubular PBR, (B) batch cylindrical PBR, (C) batch flat PBR. (D) Tornado plot showing the sensitivity analysis of various parameters on MSP, indicating the magnitude of economic influence from changes in CO_2 concentration, CO_2 cost, product yield, biomass selling price, capital cost, electricity price, and nutrient usage.

to 11 days), and reactor volumes (0.1–1000 L) (Fig. 5 A-E). Under baseline conditions of 2.5 % CO_2 concentration and long cultivation time (180 h) in helical tubular PBRs of 10*L*, the minimum selling price of lycopene was approximately $37.47/kg. Among the different combinations of these variables, the lowest production cost could reach to $2.19/kg with 5 % CO_2 concentration and 5 days of cultivation, which is comparable the market price range from $5.4 to $170/kg with high volatility (Tridge, 2025). A similar simulation was done for plastic cylindrical PBR and flat PBR in batch mode (Fig. 5B, C). Among the different combinations of these variables, the lowest production cost from cylindrical PBR could reach to $130.94/kg with 0.5 % CO_2 concentration, 1000 L reaction size, and 11 days of cultivation. Under the same cultivation conditions, the production cost of lycopene using flat PBR could reach $175/kg. This adaptability across products and reactor configurations demonstrates the versatility of MAP in guiding bioprocess design and optimization.

Then, a sensitivity analysis was conducted to identify the most influential variables for production cost estimation (Fig. 5D). Among the evaluated factors, CO_2 concentration and cost, and product yield exhibited the highest sensitivity, followed by biomass selling price and PBR capital costs. These insights underscore the importance of refining cultivation strategies and reactor design to manage cost drivers effectively. Together, these visual tools demonstrate how the AI–TEA integration not only accelerates scenario testing but also enhances transparency and decision-making in bioprocess development.

The framework of MAP enabled rapid generation of over 1,500 simulated conditions, greatly expanding the scale of TEA exploration beyond what would be feasible with manual setup. Fig. 6 A–E illustrates how cultivation time, CO_2 concentration, and reactor type interact to shape production cost across reactor scales from 0.1 L to 1000 L (simulation data of flat PBRs in batch is provided additionally). The 3D surfaces highlight differences between semi-continuous tubular reactors and batch-mode cylindrical systems, providing insight into cost trends and design trade-offs. To further visualize the interaction between AI-predicted parameters and economic outcomes, a 3D scatter plot maps the TL outputs (CO_2 concentration, product yield, and biomass yield) against the resulting minimum selling price (Fig. 6F). This plot illustrates how variations in biological performance metrics directly shape process economics, enabling intuitive exploration of parameter trade-offs and optimization opportunities. Regions associated with higher biomass productivity and growth performance correspond to lower MSP values, validating the value of predictive modeling in early-stage screening.

To test the generalizability of this AI-TEA streamline, a preliminary

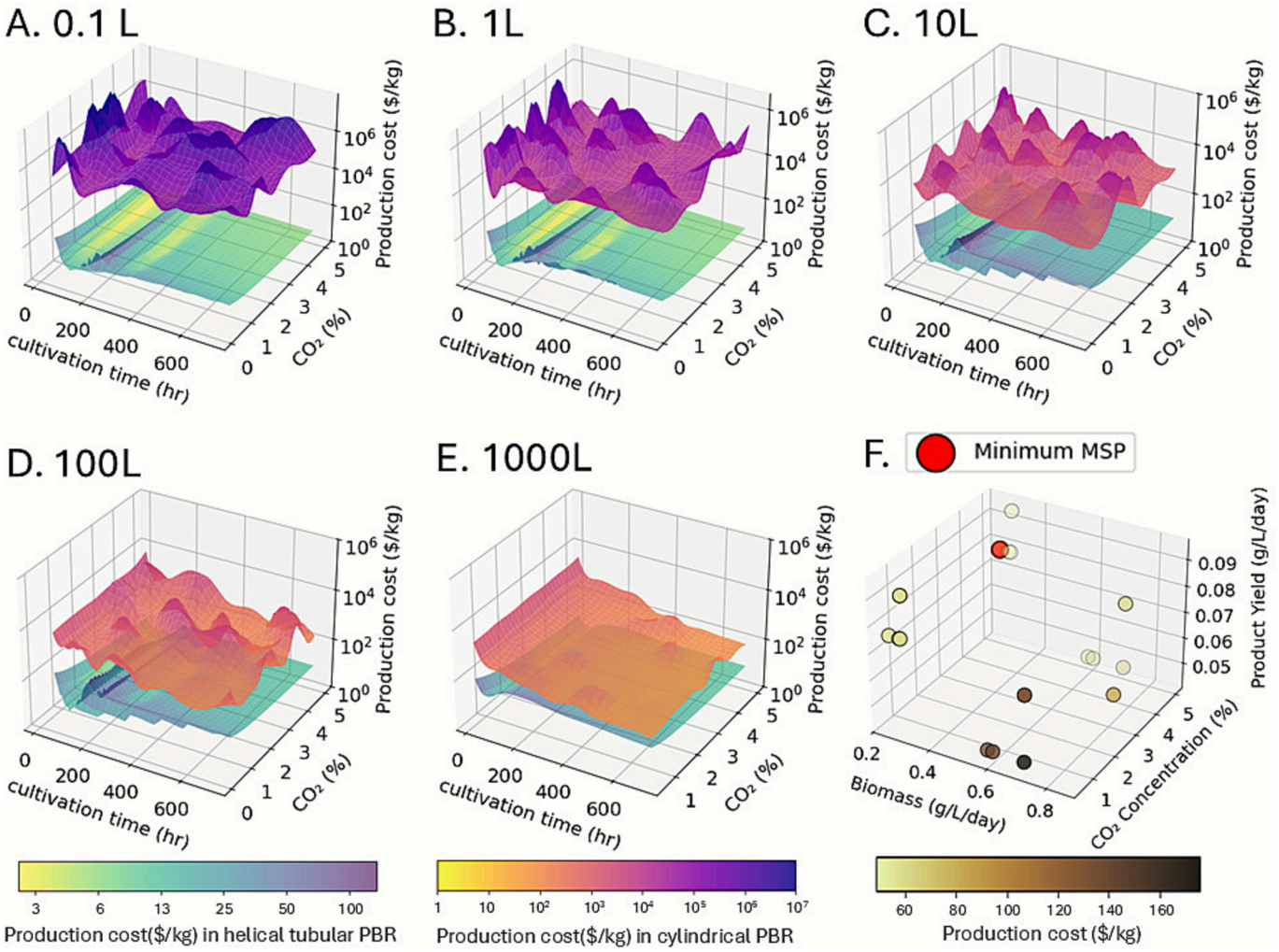


Fig. 6. Techno-economic analysis of helical tubular PBR and cylindrical PBR cases. A-E Three-dimensional surface plots illustrating the relationships between cultivation time, CO_2 concentration, and economic outcomes represented as production cost in (A) 0.1L, (B) 1 L, (C) 10 L, (D) 100 L, and (E) 1000 L PBRs. Blue-green surfaces shows production cost for lycopene using helical tubular PBR in semi-continuous mode, purple-red color shows production cost of lycopene using plastic cylindrical PBR in batch mode. (F) Three-dimensional scatter plot demonstrating the impact of biomass productivity, product yield, and CO_2 concentration on MSP, with the red marker highlighting conditions achieving the lowest MSP.

analysis for sucrose production was performed using batch-mode cylindrical reactors. The result suggested a minimum cost of $2/kg under low CO_2 (0.5 %) and extended cultivation conditions. While MAP-TEA integration facilitates rapid scenario evaluation, AI-predicted titers at the upper extreme may surpass current physiological and engineering limits, particularly for cyanobacteria under industrially relevant conditions. Known constraints, such as maximum carbon fixation rates, cell aggregation impacting light utilization, and extraction bottlenecks, often result in diminishing returns as projected titers and scales increase. Predicted values that substantially exceed experimental benchmarks should therefore be interpreted as ideal scenario bounds, not realized achievements. Future expansions will incorporate mechanistic constraints and continuous recalibration against new scale-up data to improve predictive realism.

Overall, by combining predictive modeling with economic analysis, the MAP framework enables rapid, scenario-based screening of engineered strains and cultivation strategies. It reduces reliance on early-stage lab testing and supports data-driven decisions for scale-up of photosynthetic products.

3.6. Future direction

While MAP demonstrates advanced application of multimodal AI technology to bioproduction and streamlined knowledge-to-cost analysis, there are still limitations. One limitation of the current study is reliance on literature-derived data from diverse research groups, inherently containing experimental biases, inconsistencies, and incomplete reporting, which affects model robustness and generalizability. Further validation with unpublished or pristine experimental datasets would provide an even stronger test of this framework. The underrepresentation of low-performing strains and atypical pathway engineering reduce predictive accuracy. Additionally, the current framework lacks validation from industrial datasets, limiting confidence in commercial applications. On the other hand, feature extraction using GPT-4 still requires manual curation to ensure reliability, highlighting the dependency on domain expertise. Moreover, dynamic and temporal factors influencing bioproduction were insufficiently captured, indicating a need to integrate time-series and multi-omics data to achieve more accurate predictive modeling.

Future enhancements of MAP may integrate causal inference methodologies to clarify relationships among genetic modifications, cultivation conditions, and productivity outcomes, addressing uncertainties identified in feature importance analyses. Future iterations of MAP may also benefit from comparison with recent LLM-enabled scientific information extraction systems such as PaperQA, SciBERT, and BioGPT (Lála et al., 2023; Beltagy et al., 2019; Luo et al., 2022). While these platforms facilitate automated literature understanding, they principally focus on textual question-answering or biomedical data abstraction and remain limited in multimodal process integration, specifically targeting underrepresented phenotypes like low-production strains. Additionally, incorporating multimodal sensing data, such as metabolomics and transcriptomics, could improve the accuracy of predicting metabolic bottlenecks. Furthermore, extending the MAP framework to support automated robotic experimentation platforms could enable high-throughput, closed-loop optimization workflows, accelerating strain development and scaling-up efforts for commercial photobiorefinery applications. Beyond its current static design, integrating MAP with active learning and real-time experimental feedback would create an adaptive AI system capable of self-refining knowledge boundaries. Such an extension would enable closed-loop DBTL cycles that continuously improve predictive fidelity and process optimization for photobiorefinery applications (Helmy et al., 2023; Long et al., 2022). Lastly, a promising direction is to integrate this multimodal AI application with an agentic AI system, enabling autonomous, goal-driven decision-making to accelerate innovation and sustainability in the cyanobacteria-empowered bioeconomy.

4. Conclusions

MAP integrates GPT-4 assisted data mining, TL, and TEA to model cyanobacterial bioproduction. Specifically, TL improved predictive accuracy for UTEX 2973 ($R^2 > 0.93$ under random splits; > 0.76 in stringent test), with PCC 7942 providing the strongest knowledge transfer. Cultivation features dominated growth predictions, while product and genetic features drove titer and production rate. AI-driven TEA explored over 1,500 lycopene production scenarios across reactor types, CO_2 levels, and cultivation times. This multimodal approach streamlines *in silico* DBTL cycles toward economically viable photobiorefinery development.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT to assist data extraction and improve writing language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Runyu Zhao: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wenyu Li:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jose Gonzalez-Aguirre:** Writing – original draft, Methodology, Investigation, Formal analysis. **Yayun Chen:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Joshua S. Yuan:** Writing – review & editing, Supervision, Funding acquisition. **Himadri B. Pakrasi:** Writing – review & editing, Supervision, Funding acquisition. **Chengcheng Fei:** Writing – review & editing, Supervision, Funding acquisition. **Sunkyu Park:** Writing – review & editing, Supervision, Funding acquisition. **Garrett Roell:** Writing – review & editing, Resources, Data curation. **Yixin Chen:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Yinjie J. Tang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by the B-SURE program of Defense Advanced Research Projects Agency, U.S. (HR001122S0010), Fossil Energy Office of the Department of Energy, United States (DE-FE0032108), Energy Earthshots award of Deparment of Energy, United States (DE-SC0024702), and Carbon Utilization Redesign for Biomanufacturing (CURB) Engineering Research Center (ERC) of National Science Foundation, United States (EEC 2330245). Fig. 1 created in BioRender. Zhao, R. (2025) https://BioRender.com/ww312ba.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2025.133646.

Data availability

Data will be made available on request.

References

Adomako, M., Ernst, D., Simkovsky, R., Chao, Y.Y., Wang, J., Fang, M., Bouchier, C., Lopez-Igual, R., Mazel, D., Gugger, M., Golden, S.S., 2022. Comparative Genomics of Synechococcus elongatus explains the Phenotypic Diversity of the Strains. MBio 13 (3), e0086222. https://doi.org/10.1128/mbio.00862-22.

Agarwal, P., Soni, R., Kaur, P., Madan, A., Mishra, R., Pandey, J., Singh, S., Singh, G., 2022. Cyanobacteria as a Promising Alternative for Sustainable Environment: Synthesis of Biofuel and Biodegradable Plastics. Front. Microbiol. 13–2022. https://doi.org/10.3389/fmicb.2022.939347.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable Artificial Intelligence (XAI): what we know and what is left to attain Trustworthy Artificial Intelligence. Inf. Fusion 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805.

Banerjee, S., Ramaswamy, S., 2019. Comparison of productivity and economic analysis of microalgae cultivation in open raceways and flat panel photobioreactor. Bioresour. Technol. Rep. 8, 100328. https://doi.org/10.1016/j.biteb.2019.100328.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv preprint arXiv:1903.10676. https://doi.org/10.48550/arXiv.1903.10676.

Billis, K., Billini, M., Tripp, H.J., Kyrpides, N.C., Mavromatis, K., 2014. Comparative Transcriptomics between Synechococcus PCC 7942 and Synechocystis PCC 6803 provide Insights into Mechanisms of stress Acclimation. PLoS One 9 (10), e109738. https://doi.org/10.1371/journal.pone.0109738.

Broddrick, J.T., Rubin, B.E., Welkie, D.G., Du, N., Mih, N., Diamond, S., Lee, J.J., Golden, S.S., Palsson, B.O., 2016. Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. PNAS 113 (51). https://doi.org/10.1073/pnas.1613446113.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., Collins, J.J., 2018. Next-Generation Machine Learning for Biological Networks. Cell 173 (7), 1581–1592. https://doi.org/10.1016/j.cell.2018.05.015.

Casa, M., Miccio, M., 2021. Thinking, Modeling and Assessing costs of Extracting Added-value Components from Tomato Industrial By-products on a Regional Basis. Chem. Eng. Trans. 87. https://doi.org/10.3303/CET2187028.

Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D., 2020. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 48 (D1). https://doi.org/10.1093/nar/gkz862.

Chen, L., Li, Q., Nasif, K.F.A., Xie, Y., Deng, B., Niu, S., Pouriyeh, S., Dai, Z., Chen, J., Xie, C.Y., 2024. AI-Driven Deep Learning Techniques in Protein Structure Prediction. Int. J. Mol. Sci. 25 (15). https://doi.org/10.3390/ijms25158426.

Czajka, J.J., Oyetunde, T., Tang, Y.J., 2021. Integrated knowledge mining, genome-scale modeling, and machine learning for predicting Yarrowia lipolytica bioproduction. Metab. Eng. 67, 227–236. https://doi.org/10.1016/j.ymben.2021.07.003.

Davis, R., Markham, J., Kinchin, C., Grundl, N., Tan, E.C., Humbird, D., 2016. Process design and economics for the production of algal biomass: algal biomass production in open pond systems and processing through dewatering for downstream conversion. National Renewable Energy Lab (NREL), Golden, CO (United States). https://doi.org/10.2172/1239893.

Decker, S.R., Brunecky, R., Yarbrough, J.M., Subramanian, V., 2023. Perspectives on biorefineries in microbial production of fuels and chemicals. Frontiers in Industrial Microbiology 1–2023. https://doi.org/10.3389/finmi.2023.1202269.

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., Wishart, D.S., 2019. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J. Cheminform 11 (1), 2. https://doi.org/10.1186/s13321-018-0324-5.

Eslami, M., Adler, A., Caceres, R.S., Dunn, J.G., Kelley-Loughnane, N., Varaljay, V.A., Martin, H.G., 2022. Artificial intelligence for synthetic biology. Commun. ACM 65 (5), 88–97. https://doi.org/10.1145/3500922.

Flores, J.E., Claborne, D.M., Weller, Z.D., Webb-Robertson, B.-J.-M., Waters, K.M., Bramer, L.M., 2023. Missing data in multi-omics integration: recent advances through artificial intelligence. Front. Artif. Intell. 6–2023. https://doi.org/10.3389/frai.2023.1098308.

Guo, B., Lu, X., Jiang, X., Shen, X.L., Wei, Z., Zhang, Y., 2025. Artificial Intelligence in advancing Algal Bioactive Ingredients, Production, Characterization, and Application. Foods 14 (10). https://doi.org/10.3390/foods14101783.

Helmy, M., Elhalis, H., Liu, Y., Chow, Y., Selvarajoo, K., 2023. Perspective: Multiomics and Machine Learning help Unleash the Alternative Food potential of Microalgae. Adv. Nutr. 14 (1), 1–11. https://doi.org/10.1016/j.advnut.2022.11.002.

Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., Müller, H., 2023. AI for life: Trends in artificial intelligence for biotechnology. N. Biotechnol. 74, 16–24. https://doi.org/10.1016/j.nbt.2023.02.001.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., Lewis, N.E., 2016. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 44 (D1). https://doi.org/10.1093/nar/gkv1049.

Knoot, C.J., Ungerer, J., Wangikar, P.P., Pakrasi, H.B., 2018. Cyanobacteria: Promising biocatalysts for sustainable chemical production. J. Biol. Chem. 293 (14), 5044–5052. https://doi.org/10.1074/jbc.R117.815886.

Kugler, A., Stensjö, K., Kugler, A., Stensjö, K., 2023. Optimal energy and redox metabolism in the cyanobacterium Synechocystis sp. PCC 6803. npj Syst. Biol. Appl. 9 (1). https://doi.org/10.1038/s41540-023-00307-3.

Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S.G., White, A.D., 2023. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research. arXiv preprint arXiv:2312.07559. https://doi.org/10.48550/arXiv.2312.07559.

Li, Z., Shubin, L., Lei, C., Tao, S., Zhang, W., 2024. Fast-growing cyanobacterial chassis for synthetic biology application. Crit. Rev. Biotechnol. 44 (3), 414–428. https://doi.org/10.1080/07388551.2023.2166455.

Li, W., Mao, Z., Xiao, Z., Liao, X., Koffas, M., Chen, Y., Ma, H., Tang, Y.J., 2025. Large language model for knowledge synthesis and AI-enhanced biomanufacturing. Trends Biotechnol. https://doi.org/10.1016/j.tibtech.2025.02.008.

Lobentanzer, S., Rodriguez-Mier, P., Bauer, S., Saez-Rodriguez, J., 2024. Molecular causality in the advent of foundation models. Mol. Syst. Biol. 20 (8), 848–858-858. https://doi.org/10.1038/s44320-024-00041-w.

Long, B., Fischer, B., Zeng, Y., Amerigian, Z., Li, Q., Bryant, H., Li, M., Dai, S.Y., Yuan, J. S., 2022. Machine learning-informed and synthetic biology-enabled semi-continuous algal cultivation to unleash renewable fuel productivity. Nat. Commun. 13 (1), 541. https://doi.org/10.1038/s41467-021-27665-y.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T., 2022. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. arXiv preprint arXiv:2210.10341. https://doi.org/10.48550/arXiv.2210.10341.

Martins, R., Sales, H., Pontes, R., Nunes, J., Gouveia, I., 2023. Food Wastes and Microalgae as sources of Bioactive Compounds and Pigments in a Modern Biorefinery: a review. Antioxidants 12 (2), 328. https://doi.org/10.3390/antiox12020328.

Mateu-Sanz, M., Fuenteslópez, C.V., Uribe-Gomez, J., Haugen, H.J., Pandit, A., Ginebra, M.-P., Hakimi, O., Krallinger, M., Samara, A., 2024. Redefining biomaterial biocompatibility: challenges for artificial intelligence and text mining. Trends Biotechnol. 42 (4), 402–417. https://doi.org/10.1016/j.tibtech.2023.09.015.

McLaughlin, J.A., Myers, C.J., Zundel, Z., Misirli, G., Zhang, M., Ofiteru, I.D., Goni-Moreno, A., Wipat, A., 2018. SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. ACS Synth Biol 7 (2), 682–688. https://doi.org/10.1021/acssynbio.7b00403.

Mondal, P.P., Galodha, A., Verma, V.K., Singh, V., Show, P.L., Awasthi, M.K., Lall, B., Anees, S., Pollmann, K., Jain, R., 2023. Review on machine learning-based bioprocess optimization, monitoring, and control systems. Bioresour. Technol. 370, 128523. https://doi.org/10.1016/j.biortech.2022.128523.

Mueller, T.J., Ungerer, J.L., Pakrasi, H.B., Maranas, C.D., 2017. Identifying the Metabolic differences of a Fast-Growth Phenotype in Synechococcus UTEX 2973. Sci. Rep. 7 (1). https://doi.org/10.1038/srep41569.

Nogales, J., Gudmundsson, S., Knight, E.M., Palsson, B.O., Thiele, I., 2012. Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis - PubMed. PNAS 109 (7). https://doi.org/10.1073/pnas.1117907109.

Novoveská, L., Nielsen, S.L., Eroldoğan, O.T., Haznedaroglu, B.Z., Rinkevich, B., Fazi, S., Robbens, J., Vasquez, M., Einarsson, H., 2023. Overview and challenges of Large-Scale Cultivation of Photosynthetic Microalgae and Cyanobacteria. In: Marine Drugs, 21. https://doi.org/10.3390/md21080445.

Orth, J., Thiele, I., Palsson, B., 2010. What is flux balance analysis? Nat Biotechnol 28, 245–248. https://doi.org/10.1038/nbt.1614.

Pei, L., Garfinkel, M., Schmidt, M., 2022. Bottlenecks and opportunities for synthetic biology biosafety standards. Nat. Commun. 13 (1), 2175. https://doi.org/10.1038/s41467-022-29889-y.

Radivojević, T., Costello, Z., Workman, K., Garcia Martin, H., 2020. A machine learning Automated Recommendation Tool for synthetic biology. Nat. Commun. 11 (1), 4879. https://doi.org/10.1038/s41467-020-18008-4.

Santos-Merino, M., Singh, A.K., Ducat, D.C., 2019. New applications of Synthetic Biology Tools for Cyanobacterial Metabolic Engineering. Front. Bioeng. Biotechnol. 7–2019. https://doi.org/10.3389/fbioe.2019.00033.

Shah, M., Wever, M., Espig, M., 2025. A Framework for Assessing the potential of Artificial Intelligence in the Circular Bioeconomy. Sustainability 17. https://doi.org/10.3390/su17083535.

Stork, T., Michel, K.P., Pistorius, E.K., Dietz, K.J., 2005. Bioinformatic analysis of the genomes of the cyanobacteria Synechocystis sp. PCC 6803 and Synechococcus elongatus PCC 7942 for the presence of peroxiredoxins and their transcript regulation under stress. J. Exp. Bot. 56 (422), 3193–3206. https://doi.org/10.1093/jxb/eri316.

Sun, T., Li, Z., Li, S., Chen, L., Zhang, W., 2023. Exploring and validating key factors limiting cyanobacteria-based CO2 bioconversion: Case study to maximize myo-inositol biosynthesis. Chem. Eng. J. 452, 139158. https://doi.org/10.1016/j.cej.2022.139158.

Tan, C., Xu, P., Tao, F., 2022. Carbon-negative synthetic biology: challenges and emerging trends of cyanobacterial technology. Trends Biotechnol. 40 (12), 1488–1502. https://doi.org/10.1016/j.tibtech.2022.09.012.

Tridge, 2025. Global lycopene price. https://dir.tridge.com/prices/lycopene.

Ungerer, J., Lin, P.-C., Chen, H.-Y., Pakrasi Himadri, B., 2018. Adjustments to Photosystem Stoichiometry and Electron transfer Proteins are Key to the Remarkably Fast growth of the Cyanobacterium Synechococcus elongatus UTEX 2973. MBio 9 (1). https://doi.org/10.1128/mbio.02327-17.

Xiao, Z., Li, W., Moon, H., Roell, G.W., Chen, Y., Tang, Y.J., 2023. Generative Artificial Intelligence GPT-4 Accelerates Knowledge Mining and Machine Learning for Synthetic Biology. ACS Synth. Biol. 12 (10), 2973–2982. https://doi.org/10.1021/acssynbio.3c00310.

Yadav, R.D., Dhamole, P.B., 2024. Techno economic and life cycle assessment of lycopene production from tomato peels using different extraction methods. Biomass Convers. Biorefin. 14 (20), 25495–25511. https://doi.org/10.1007/s13399-023-04676-x.

Yu, J., Liberton, M., Cliften, P.F., Head, R.D., Jacobs, J.M., Smith, R.D., Koppenaal, D.W., Brand, J.J., Pakrasi, H.B., 2015. Synechococcus elongatus UTEX 2973, a fast growing cyanobacterial chassis for biosynthesis using light and CO2. Sci. Rep. 5 (1), 8132. https://doi.org/10.1038/srep08132.

Zhang, C., Tsoi, R., You, L., 2016. Addressing biological uncertainties in engineering gene circuits. Integr Biol (camb) 8 (4), 456–464. https://doi.org/10.1039/c5ib00275c.

Zhang, Y., Li, Q., Wang, J., Chang, X., Chen, L., Liu, X., 2025. Causal network inference based on cross-validation predictability. Communications Physics 8 (1), 173. https://doi.org/10.1038/s42005-025-02091-4.

Zhao, R., Sengupta, A., Tan, A.X., Whelan, R., Pinkerton, T., Menasalvas, J., Eng, T., Mukhopadhyay, A., Jun, Y.S., Pakrasi, H.B., Tang, Y.J., 2022. Photobiological production of high-value pigments via compartmentalized co-cultures using Ca-alginate hydrogels. Sci. Rep. 12 (1), 22163. https://doi.org/10.1038/s41598-022-26437-y.

Zhu, Y., Jones, S.B., Anderson, D.B., 2018. Algae farm cost model: Considerations for photobioreactors. Pacific Northwest National Laboratory (PNNL), Richland, WA, United States. https://doi.org/10.2172/1485133.