ACS Synthetic Biology

pubs.acs.org/synthbio Research Article

Generative Artificial Intelligence GPT-4 Accelerates Knowledge Mining and Machine Learning for Synthetic Biology

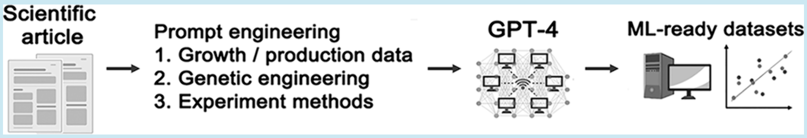
Published as part of the ACS Synthetic Biology virtual special issue "AI for Synthetic Biology".

Zhengyang Xiao, Wenyu Li, Hannah Moon, Garrett W. Roell,* Yixin Chen,* and Yinjie J. Tang*

Cite This: *ACS Synth. Biol.* 2023, 12, 2973−2982

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Knowledge mining from synthetic biology journal articles for machine learning (ML) applications is a labor-intensive process. The development of natural language processing (NLP) tools, such as GPT-4, can accelerate the extraction of published information related to microbial performance under complex strain engineering and bioreactor conditions. As a proof of concept, we proposed prompt engineering for a GPT-4 workflow pipeline to extract knowledge from 176 publications on two oleaginous yeasts (*Yarrowia lipolytica* and *Rhodosporidium toruloides*). After human intervention, the pipeline obtained a total of 2037 data instances. The structured data sets and feature selections enabled ML approaches (e.g., a random forest model) to predict *Yarrowia* fermentation titers with decent accuracy (R^2 of 0.86 for unseen test data). Via transfer learning, the trained model could assess the production potential of the engineered nonconventional yeast, *R. toruloides*, for which there are fewer published reports. This work demonstrated the potential of generative artificial intelligence to streamline information extraction from research articles, thereby facilitating fermentation predictions and biomanufacturing development.

KEYWORDS: *feature selection, natural language processing, human intervention, prompt engineering, transfer learning, Yarrowia lipolytica*

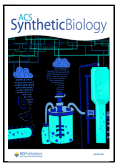
INTRODUCTION

Synthetic biology (SynBio) tools can engineer microbes for sustainable biomanufacturing. To develop microbial workhorses, researchers rely on trial and error for breakthroughs due to the complex nature of biological systems. Model predictions of cell performance are key to reducing the number of experimental trials and improving the strain development effectiveness. However, mechanistic models have difficulty incorporating all influential factors to simulate microbial production.[1] On the other hand, machine learning (ML) has been applied to predict fermentation titers,[2−4] optimize bioprocesses,[5−7] and recommend engineering approaches.[8,9] The drawback of ML is that it requires large sets of experimental data for model training. Therefore, knowledge mining from published journal articles can be an inexpensive strategy for training ML models. However, manually extracting data from a large number of articles is labor-intensive and prone to human errors and inconsistencies in quality, because reported data often lack a standardized format,[10,11] and substantial efforts are needed to interpret information and organize it into ML-ready data.[12]

NLP, a branch of AI, can process text at a large scale, enabling topic organization in published articles.[13] It has also been utilized to track adverse drug events from electronic health record notes.[14] A recent tipping point in the field of NLP was the release of GPT-4,[15] which shows 'sparks' of artificial general intelligence[16] to rapidly parse text based on user-provided context.[15] Leveraging GPT-4, relevant bioprocess features and outcomes from published papers can be extracted for rapid database growth. Moreover, GPT-4 can provide useful biomanufacturing guidelines, but its prediction of production titers from nonmodel yeasts may give unrealistic answers (Supplementary Figure 1). Therefore, this study aims to integrate GPT-4 with ML to improve the prediction of yeast fermentation titers.[2,17]

As a proof of concept, this study used GPT-4 to extract knowledge from articles on the industrial yeast *Yarrowia lipolytica*. Under human supervision, the published information was transformed into data samples (i.e., instances). Each instance included both outputs (product titers) and inputs (i.e., features). Feature variables included bioprocess con-

Received: May 17, 2023
Published: September 8, 2023



https://doi.org/10.1021/acssynbio.3c00310

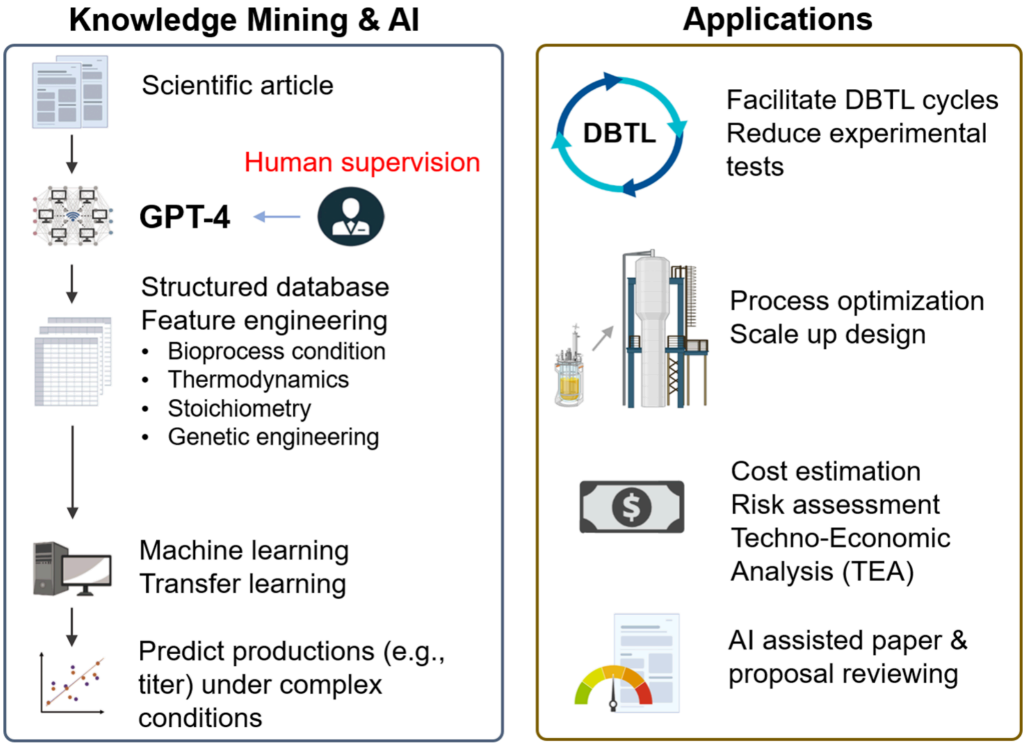


Figure 1. GPT-4 knowledge mining for ML (Left) and AI applications (Right: assist biomanufacturing design, commercial decision, or project quality/risk assessment).

ditions, metabolic pathways, and genetic engineering methods. All instances have been uploaded to a database for training ML models. Moreover, *Rhodosporidium toruloides* is a novel yeast that has recently gained research attention for its high lipid content[18,19] and native carotenoid production.[20,21] However, the literature on *Rhodosporidium* is sparse.[22] Here, we demonstrate that transfer learning can use knowledge from well-studied domains (*Yarrowia*-trained model) to understand less-studied scenarios and speed up the learning process.[23,24] In summary, for the first time, this study integrated the GPT with knowledge engineering and ML for predicting microbial cell factories. The lessons will improve human supervision and prompt engineering, facilitating GPT and ML applications in SynBio fields (Figure 1).

RESULTS AND DISCUSSION

Extraction of ML Features and Data Sets from SynBio Papers via GPT-4 Workflow. ML approaches require a large amount of experimental data to correlate ML inputs (features) with outputs (productions). Since the biomanufacturing literature presents a wealth of strain construction and bioprocess engineering case studies, constructing a database from published papers may broadly support ML applications. Previous databases, like LASER,[10] collected metabolic engineering reports, but the stored information was not organized and transformed for ML applications. In contrast, this study performed knowledge mining and feature selection, which could filter out erroneous/redundant information and capture factors that independently affect bioproduction. Moreover, SynBio papers describe bioreactor conditions, metabolic pathways, and genetic engineering methods. Manual information extraction is time-consuming. Here, GPT-4 was used to overcome this challenge. Since GPT-4's maximum context window had 8,192 tokens, the sections of each scientific article, including abstract, materials and methods, results, and data tables, were manually separated into text files. Prompts (questions for GPT) were then added to the beginning of each section (Table 1) so that GPT-4 could summarize the experiment results and methods into accessible tables (See examples in Supplementary File 1, **Screenshots**).

Quality Tests for Data Extraction by ChatGPT. Data extraction from publications without losing important knowledge is challenging. To test GPT applicability, we started to use GPT-3.5 API on March 15, 2023, to extract *Rhodosporidium* fermentation data from journal articles. The output of one PhD student using the GPT-enhanced workflow for 1 week is illustrated (Figure 2a). When GPT-3.5 was used, on average, 11.7 papers were extracted per day (8 work hours). After the release of GPT-4 on 3/15, 25 papers were extracted on 3/16. We tested the correctness of biomanufacturing data extracted by GPT-3.5 in 10 *Yarrowia* papers. Via manual examination, we found that the titer data extracted by GPT-3.5 were 74% correct. Some erroneous data were obvious as they included numbers that were consecutive, repeated, or not present in the article. With user discretion to fix the outputs, the extracted titer data were ~89% correct. When GPT-4 was applied, the quality of the extracted data was significantly improved. For example, GPT-4 accurately obtained fermentation titers from 10 *Yarrowia* papers (Figure 2b).

GPT-4 can retrieve experimental methods, engineering strategies, and fermentation outcomes. Supplementary Table 1 summarizes the data location, output format, and correctness. Although human supervision is a necessity to ensure the quality of knowledge mining, our data extraction workflow is an improvement over manual reading because: (1) it does not rely on the expertise of a single person and can be parallelized across a team, (2) it does not require high levels of effort for data recording, (3) it extracts data reproducibly, and the extracted data can be checked for errors in a *targeted* manner rather than laboriously reading each section,[25] and (4) it is adaptable to automation once the GPT-4's Application Programming Interface (API) is available.

GPT-4 Assisted Database Construction for *Y. lipolytica* Biomanufacturing. *Yarrowia lipolytica* is an industrially important yeast for bioproductions.[26] Our previous study collected information manually from ∼100 *Yarrowia* SynBio papers,[2] which took a well-trained graduate student over 400 working hours. In this study, the GTP-4 workflow was used to obtain ∼1670 additional data instances from 115 *Yarrowia* papers within 40 working hours. To facilitate the conversion of extracted information into ML inputs, we developed two supporting tables. First, a feature table (Table 2) defined the ML features and the rules of feature selection.[2] Second, we developed a ***molecule inventory*** (Supplementary File 2) to convert certain reported information into numerical or categorical feature variables. For instance, once GPT-4 identified the carbon source as glucose, this inventory table would assign feature variables related to glucose, including substrate's category and heat of combustion. Both the feature table and the molecule inventory will be continuously updated to include broader biomanufacturing features.

For further validating the applicability of GPT-4, the manually extracted data were compared with GPT-extracted data by computing feature importance, feature variances, and principal component analysis (PCA). The GPT-extracted data had a distribution of feature importance similar to that of the manually extracted data (Figure 3a), suggesting that the newly extracted data followed patterns similar to those of the manually extracted data. Interestingly, the GPT data sets had higher feature variances than the manually extracted data set for 19 of 28 features (Figure 3b). PCA showed that, with a similar silhouette score after K-means converged to the optimal solution, the data extracted by GPT had 7% higher mean distance between clusters (Figure 4). The PCA loadings further indicated that the clustering of the manually extracted data set was governed by the carbon source and product cofactor cost (Supplementary Figure 2). In contrast, GPT data were clustered according to culture condition and genetic engineering features in addition to the carbon source and cofactor cost. Therefore, GPT-4 can capture more distinctiveness within papers and reason through complex contextual data to generate less biased biomanufacturing instances.

Leveraging the GPT-Constructed Database To Predict *Y. lipolytica* Fermentation Titer. Fermentation titer determines the bioprocess economy. The GPT-assisted database construction can support the quantitative prediction of yeast fermentation titers under various conditions. Specifically, *Y. lipolytica* fermentation instances formed a comprehensive database to train ML models. We conducted a comparative test of seven classical ML algorithms with data scaling (Supplementary Figure 3, support vector machine (SVM), Gaussian process (GP), multilayer perceptron (MLP), random forest (RF), extreme gradient boosting (XGBoost), k-nearest neighbors (KNN), and linear regression). Based on unseen testing data and ML prediction (predicted titer vs reported titer), linear regression and linear SVM performed poorly, suggesting that a linear relationship cannot accurately represent the titer prediction. A fully connected two-layer neural network did not exhibit a good performance either. In contrast, the RF model achieved the best accuracy: R^2 of 0.86 on unseen test instances (Figure 5a). After this point, train/test

Table 1. Prompts Used for GPT-4 Data Extraction[a]

<table>
<tr><th>prompt number</th><th>information extracted</th><th>section</th><th>prompt text (example)</th></tr>
<tr><td>1</td><td>substrate and product; growth condition</td><td>abstract, results</td><td>a multicolumn spreadsheet from the given text| experiment number | strain name | product | product concentration | substrate | substrate concentration | media | time | mode of bioreactor operation |+ [Abstract text] + [Result text]</td></tr>
<tr><td>2</td><td>genetic engineering</td><td>materials and methods, results</td><td>a multicolumn spreadsheet from the given text | strain name | parent strain | knocked out genes | expressed genes | promoter | genome integration Y/N | codon optimized | When outputting the daughter strain, you need to include the expressed and knockout genes in the parent strain. + [Method text] + [Result text]</td></tr>
<tr><td>3</td><td>growth condition</td><td>materials and methods</td><td>a multicolumn spreadsheet from the given text | experiment type | culture volume | culture vessel | substrate | substrate concentration | media | nitrogen condition | time | pH | temperature | + [Method text]</td></tr>
<tr><td>4</td><td>tabulated experiment data</td><td>tables</td><td>convert these into spreadsheets: [Table text]</td></tr>
</table>

[a] Note: The redundancy in the information provided by each prompt is necessary to align the bioprocess data in each instance and to compile it into a final ML-ready data table.

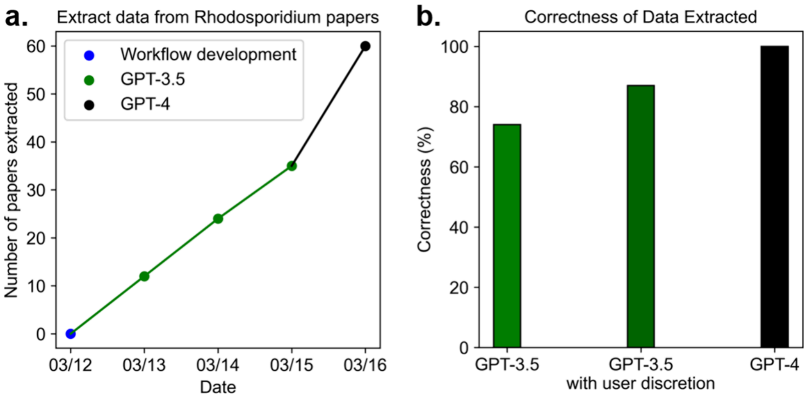


Figure 2. Data extraction effectiveness of GPT-3.5 and GPT-4. (a) Number of *Rhodosporidium* papers processed in 5 days by a single user. (b) Correctness of fermentation data extracted from a test set of 10 *Yarrowia* articles (note: extracted data were manually inspected).

data were not scaled; therefore, their original physical meanings were retained. The RF model still showed robust performance for test data with an average R^2 of 0.80 ± 0.04 across 50 random data splits. The test set performance of the RF regressor was decent for nearly all product classes: organic acid, lipid, terpene, flavonoid, fatty-acid-derived compound, sugar alcohol, glycan, and polyketide (Figure 5b−k). The new ML model, trained on a database approximately 50% larger than the previous model,[2] showed general improvements for titer predictions of small terpene and polyketide products. However, the ML model was still unable to explain the data related to large terpenes, as indicated by the region with horizontal consecutive dots in Figure 5e. This is because some key biomanufacturing features are still missing from knowledge mining. For instance, a recent study removed lycopene substrate inhibition by site-mutagenesis enzyme engineering, and it was able to achieve the highest β-carotene titer ever reported, 39.5 g/L.[27] Besides, another research showed that fermentation yields were improved through cytoplasmic-peroxisomal compartmentalization engineering.[28−30] The DNA sequences of key genes also affect pathway performances.[31,32] However, the current database cannot obtain these influential factors at the molecular level.

Transfer Learning From *Y. lipolytica* to *R. toruloides* To Reveal the Genetic Engineering Effect on Production Performance. Nonmodel cell factories have been rapidly developing. For example, *R. toruloides* is a nonmodel yeast that can convert cheap feedstock into high-value carotenoids. However, this species has received few genetic engineering reports. Transfer learning could leverage knowledge from the Yarrowia data set to reveal the potential genetic engineering outcomes. For instance, we extracted 366 *Rhodosporidium* fermentation results from 60 articles to train the RF model for predicting lipid and biomass production (R^2 > 0.4) (Supplementary Figure 4), but the database lacks genetic engineering features. For instance, reports on astaxanthin production in *R. toruloides* mainly focused on its native pathways.[33−35] Therefore, knowledge transferring from *Y. lipolytica* reports is necessary to predict how genetic engineering affects astaxanthin production in *R. toruloides* and offers guidelines for future strain development. Here, we utilized two inductive learning approaches[36]: (1) a neural network with a pretrained encoder-decoder structure to study the effect of the number of gene expressions on the astaxanthin synthesis; (2) an instance-based random forest TL approach to address the source-target domain gap. We evaluated these two approaches because they have different underlying assumptions. The encoder−decoder structure is a type of neural network that transforms the embedded data to a latent space, takes this transform representation, and attempts to reconstruct the original data set. A pretrained encoder implies that *Yarrowia* and *Rhodosporidium* data are from the same knowledge domain and have the same statistical distribution. In contrast, the instance-based random forest model can handle data from two different knowledge domains.

First, a pretrained encoder in an autoencoder[37] was used to reduce the number of features from 29 (original 28 features +1 categorical input of species, *Yarrowia* or *Rhodosporidium*) to 14. The resulting model predicted *R. toruloides* astaxanthin production after 96 h of shake flask cultures with rich media. However, the titer output from the trained model was insensitive to genetic modification features (Supplementary Figure 5). In the reported experiments, the titer for *R. toruloides* astaxanthin was ∼1 mg/L. In contrast, the majority of instances in the ML database for *Y. lipolytica* and *R. toruloides* productions were at the g/L level. The order of magnitude difference between astaxanthin and other products (e.g., lipids and biomass training data) made predicting low-titer products difficult. Besides, encoding input data might impair the neural network's ability to learn biological features from complex and interconnected biological systems.

Second, a RF transfer learning was tested. This approach combined *Y. lipolytica* engineering experiences with the reported growth characteristics of *R. toruloides*, which predicts *R. toruloides* biomanufacturing potentials once its genetic tools are developed. Specifically, the training instances were labeled as either *Yarrowia* or *Rhodosporidium* (categorical input feature), and a weight of 3× was assigned to the *Rhodosporidium* data. The model trained on data from both species was used to predict *R. toruloides* astaxanthin titers. Again, the inputs corresponded to a 96 h fermentation in shake flasks containing rich media. The model predicted that wild-type *R. toruloides* without genetic engineering would likely produce an astaxanthin titer below 4.2 mg/L after process optimizations (Figure 6a). This result is comparable with a recent publication (not used in the model training) reporting astaxanthin production of 1.3 mg/L by *R. toruloides* in shake flasks.[35] With successful gene expressions, the engineered strains were predicted to increase their titers (Figure 6b−d). A strain might achieve an average of 39.5 mg/L astaxanthin if six key genes could be optimized (Figure 6d). Moreover, the broad distribution of astaxanthin production observed in the

Table 2. Conversion of GPT Information Into Biomanufacturing Features[a]

no	feature type	feature name	data type	description
1	substrates and products	carbon source 1	categorical	the primary carbon source used in the yeast growth media. for example, we assigned categorical numbers: glucose as 1, glycerol as 2, citrate as 3, and so on.
2		carbon source 1 concentration, g/L	numeric	the concentration of primary carbon source.
3		carbon source 1 heat of combustion, kJ/mol	numeric	the energy content of primary carbon source.
4		carbon source 1 heat of combustion, kJ/g	numeric	the energy content of primary carbon source when complex feedstock (organic waste, oil, or whey) was used.
5		carbon source 2	categorical	the secondary carbon source used in the growth media.
6		carbon source 2 concentration, g/L	numeric	the concentration of secondary carbon source.
7		carbon source 2 heat of combustion, kJ/mol	numeric	the energy content of secondary carbon source.
8		carbon source 2 heat of combustion, kJ/g	numeric	the energy content of secondary carbon source when complex feedstock (organic waste, oil, or whey) was used.
9		product class	categorical	10 product categories: organic acid, lipid, small terpene, large terpene, flavonoid, fatty acid-derived, sugar alcohol, glycan, polyketide, and biomass.
10		product Gibbs energy of formation, kJ/mol	numeric	the thermodynamic energy barrier to form the product.
11	stoichiometry	number of carbon atoms in the product	numeric	number of carbon atoms in one product molecule.
12		number of hydrogen atoms in the product	numeric	number of hydrogen atoms in one product molecule.
13		number of oxygen atoms in the product	numeric	number of oxygen atoms in one product molecule.
14		M_w of the product	numeric	the molecular weight of the product.
15	metabolic pathway	number of pathway enzymatic steps	numeric	the number of enzymatic steps to form the product counting from the closest central metabolic precursor.
16		ATP cost per product molecule	numeric	the number of ATP required to form one product molecule from the closest central metabolic pathway precursor.
17		NADH/NADPH cost per product molecule	numeric	reducing equivalence required to form one product molecule from the closest central metabolic pathway precursor.
18	growth and bioreactor condition	fermentation time, h	numeric	total time of yeast growth.
19		reactor type	nategorical	shake flask = 1, batch reactor = 2, chemostat = 3, fed batch = 4
20		culture volume, L	numeric	working volume of the cultivation vessel.
21		medium type	categorical	the type of medium used (e.g., rich medium, defined medium and minimal medium).
22		temperature, °C	numeric	cultivation temperature.
23		pH	numeric	cultivation pH.
24		nitrogen sources	categorical	nitrogen content in the growth media (on a scale of 1−3, 3 is the most sufficient).
25	genetic engineering	number of genes modified	numeric	the total number of genes modified, including deletion of native genes, overexpression of native genes, and expression of heterologous genes.
26		number of genes deleted	numeric	the number of genes knockout.
27		number of native genes overexpressed	numeric	the number of native genes overexpressed.
28		number of heterologous genes	numeric	the number of heterologous genes expressed.

[a] Note: Numbers 1, 2, 5, 6, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, and 28 are directly from GPT-4 output. Numbers 3, 4, 7, 8, 10, 11, 12, 13, 14, 15, 16, and 17 require molecule inventory for converting GPT information into feature variables. Numbers 24, 26, 27, and 28 require manual correction. If one feature was not stated in the article, it was left blank in the data set. Supplementary File 2 describes categorical number assignment.

bootstrapping analysis suggests that there is a significant level of uncertainty when predicting titers. In summary, RF with an instance-transfer method can give reasonable generalization predictions when the database is incomplete.

GPT-4's Limitations and Future Directions for SynBio Studies. This study has several limitations. First, GPT-4 is unable to extract graphical data. In the future, multimodal language models will be necessary to interpret images and figures.[15,38] Second, GPT-4 would stop after several rows due to token limitations (around 8k), and the context length is a bottleneck for large-scale GPT applications. Third, during our data extraction process, we encountered instances when GPT-4 could not differentiate between promoters and genes or mixed up native and heterologous genes. Without knowledge of GPT-4's self-supervised learning mechanism, it is difficult to explain its performance because of its nondeterministic nature (multiple potential outputs for identical input prompts). This problem is common to generative AIs because of their complex network parameters during training. Besides, GPT occasionally provides seemingly plausible yet nonfactual answers because of misunderstanding the prompt. To resolve these problems, few-shot prompt engineering, human intervention, and reinforcement learning through human feedback should be developed.[15,39] Fourth, feature selection in this study is still insufficient and suboptimal. For example, *Y. lipolytica* terpenoid titer was typically achieved by employing heterologous enzymes with high activities.[40,41] However, our database lacks a quantitative knowledge of enzyme activities. Finally,

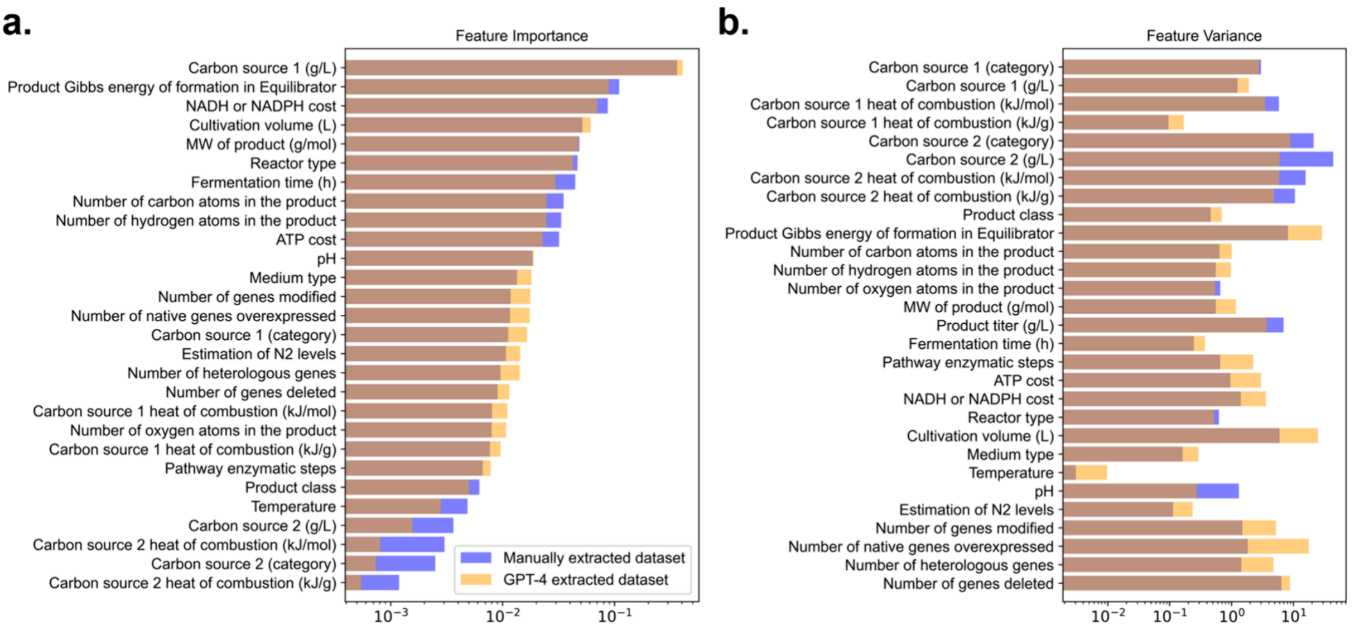


Figure 3. Comparison of the manually extracted *Yarrowia* data set with the GPT-4 extracted *Yarrowia* data set. (a) Feature importance determined by using a random forest regressor, ranked from high to low. (b) Normalized feature variance. Legend: In our visualizations, the manually extracted data set is purple, the GPT-4 extracted data set is yellow, and their overlap appears brown.

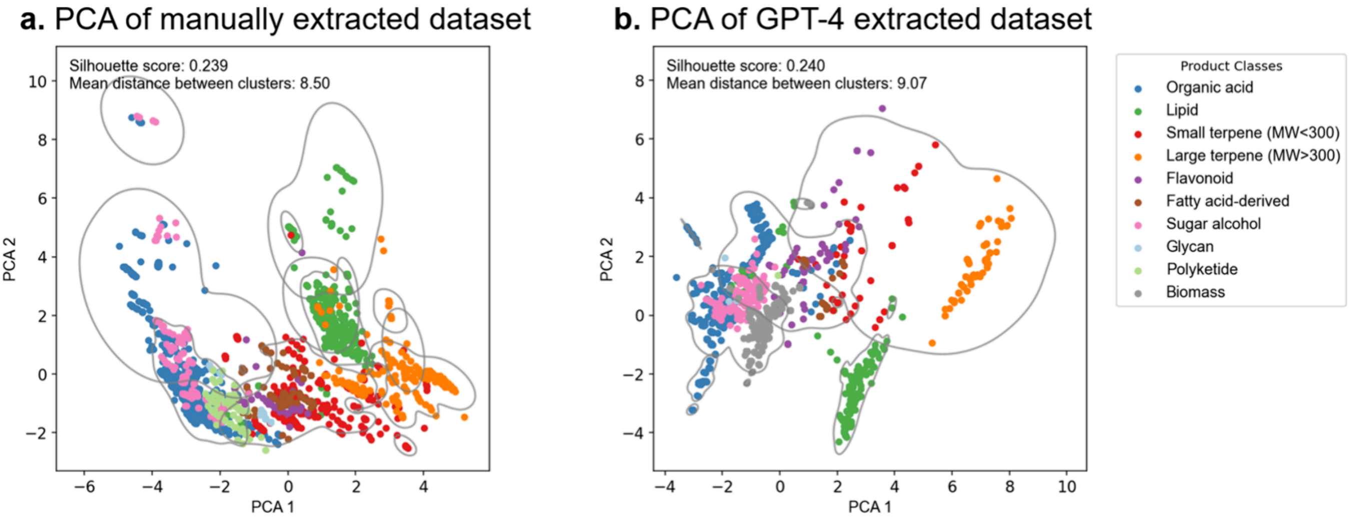


Figure 4. PCA using K-means unsupervised learning. (a) PCA of the manually extracted data set. (b) PCA of GPT-4 extracted data set. Note the axis scale difference between panels a and b.

our ML model does not use metabolic flux features. By integration with genome-scale models, ML may combine with flux balance analysis for computational strain design.[2]

CONCLUDING REMARKS

Our work aims to leverage the power of GPT to automate knowledge mining from the existing literature for supporting ML applications. Here, GPT-4 can process vast amounts of information, reducing the effort researchers need to spend on literature analysis. Particularly, model microbial hosts, such as *S. cerevisiae* and *E. coli*, have tens of thousands of relevant articles. There are great opportunities to use GPT to revolutionize biomanufacturing data science and implement ML/transfer learning to accelerate design-build-test-learn for microbial factory development.

METHODS

Data Extraction From Journal Articles. The GPT-3.5 version[39] used was between the dates 3/10 to 3/15/2023. The GPT-4 version was between 3/16 and 4/2/2023. Data extraction and feature organizations were done in a semi-automated fashion because GPT-4 was only available through OpenAI's website. The workflow can be summarized as follows: Label article sections + enter prompts → Record GPT response → Manual quality check → Convert the information into ML-ready data set based on molecule inventory and feature table

First, text files for the article sections (title, abstract, method, results, and text-based tables) were labeled and then input into GPT along with the corresponding prompt sentence (Table 1, the prompt sentences were designed based on our previous study).[2] Second, GPT-4's responses were recorded, and the

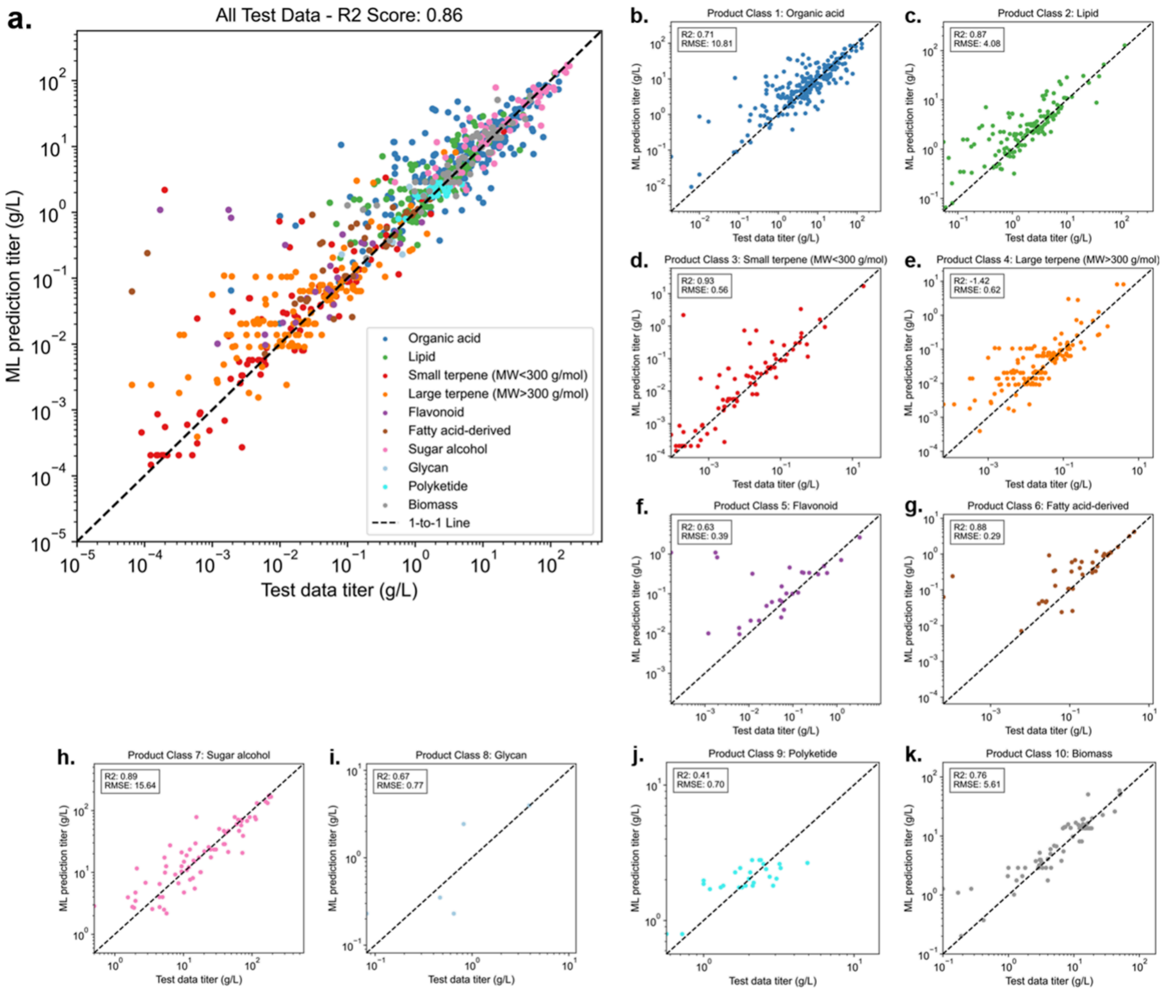


Figure 5. *Y. lipolytica* titer predictions in the test dataset using a random forest ensemble learner.

data correctness was checked manually. Third, the extracted information was transformed into an ML-ready format via the feature table. The extracted data and study citations were deposited in both Supplementary File 2 and an online database ImpactDB (https://impact-database.com/, currently under development). To build the molecule inventory, thermodynamics data for substrates and products were obtained from the NIST webbook (https://webbook.nist.gov/chemistry/) or eQuilibrator 3.0 (https://equilibrator.weizmann.ac.il/). Pathway enzyme steps, ATP, and NAD(P)H costs were estimated either by consulting KEGG[42] (https://www.genome.jp/kegg/) or reading the pathway map in relevant journal articles.

Manual and AI-Extracted Data Analysis, Data Preprocessing, and ML. All classical ML algorithms were implemented using the scikit-learn Python package. The parameters chosen for the ML model comparative test were tabulated in Supplementary Table 2. For the calculation of feature variances in comparing manually extracted data and AI-extracted data, features were normalized with their mean values and then the variances were calculated. The feature importance was determined by training an RF regressor and extracting the corresponding weights. The clustering of PCA space was done by K-means unsupervised learning using a cluster number of 6, and Euclidean distance was used to determine the distance between cluster centroids. Data preprocessing for both ML and TL are specified in the following steps. A simple imputation was performed, with missing values as zero to maintain its biological meaning. A stratified data split was performed to ensure that at least 20% data of each product class were included in the testing data. Categorical data were encoded using ordinal encoding, enabling feature comparison after training and testing.

Transfer Learning via Pretrained-encoder and RF with Instance-transfer. Transfer learning leverages knowledge gained from solving one task and applies it to improve the performance of a related but distinct task. It uses the learned representations or knowledge acquired during the training of a pre-existing model, often referred to as the "source" task, and applies this knowledge to a different, "target" task. In this work, we adopted two approaches: fine-tuning a pretrained encoder−decoder and instance transfer on RF. When tuning the pretrained encoder−decoder, we used a slow learning rate, so the model updated on the new learning task while retaining knowledge from the previous task. Conversely, instance transfer primarily combines the source and target data sets. We first trained the autoencoder structure on all *Y. lipolytica* as source data, then transferred the encoder by freezing the layers to retrain it on *R. toruloides* data without astaxanthin data. The resulting embedding was then used to test the RF model. The best result for predicting different numbers of heterologous genes used the first encoder-decoder model in Table 3. We implemented three distinct encoder−decoder structures to predict the product titer. Each structure varied in terms of the loss functions, their overall structure, and the methods

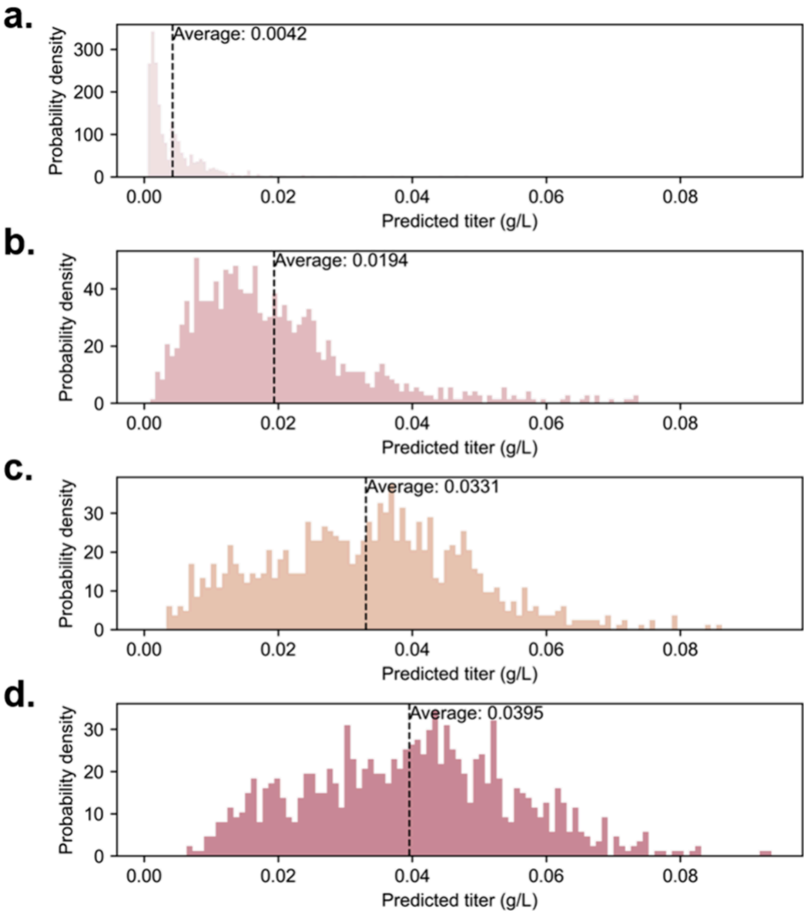


Figure 6. TL prediction for astaxanthin production in *R. toruloides*. (a) Zero heterologous gene expressed. (b) Two heterologous genes expressed. (c) Four heterologous genes expressed. (d) Six heterologous genes expressed.

Table 3. Details of the Encoder−Decoder Structure

	loss equation	model structure	latent space dimension reduction
1	L_tot = L_Reconst + l1_weight × L_l1	2-layer encoder +2-layer decoder with normalization	50% reduction
2	L_tot = triplet loss + l1_weight × l1_penalty		50% reduction
3	L_tot = L_Reconst + l1_weight × L_l1		0% reduction

employed for dimensionality reduction. Note that all layers were fully connected, and early stopping was applied in both pretraining and retraining to prevent overfitting due to noise and outliers. The model architectures were summarized in Table 3 and illustrated in Supplementary Figure 6. The loss equation is a quality measure of the model. Reconstruction loss was calculated as the MSE between the original x and embedded x, and the l1 loss (absolute error loss) was the regression score calculated after encoding. Triplet loss was calculated as the Euclidean distances between the negative sample/positive sample and the anchor, then applying the Rectified Linear Unit on the difference between the distances added by the margin. The weights of the neurons with ReLU activation were initialized using 'He' initialization[37] for all experiments. Latent space dimension reduction was the percentage of features reduced from the original data set.

The resulting performance metrics for the regression task to predict astaxanthin product titers are shown in Supplemental File 1.

The corresponding 0, 2, and 4 heterologous gene expressions samples were encoded as input data, and the prediction was performed by training a separate RF on a synthetic data set obtained by bagging with replacement 100 times on all astaxanthin data since the sample size for astaxanthin is small. The RF instance transfer was performed by the following steps. To balance the *R. toruloides*' effect on the prediction due to its small size, we augmented the *R. toruloides* data to three times its initial size to form the combined data set. The same data preprocessing was followed. The model was then tuned on the new training set, and prediction is done using Astaxanthin samples with 0, 2, 4, and 6 heterologous genes expressed respectively. The trained model was used to make predictions about the product titer. All ML and TL codes are available at https://github.com/wenyuli23/SyntheticBiologyTL.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.3c00310.

GPT-4's response screenshots and additional details of the machine learning workflow (PDF).

GPT-4-extracted *Y. lipolytica* and *R. toruloides* bioproduction database (XLSX).

AUTHOR INFORMATION

Corresponding Authors

Garrett W. Roell − *ImpactDB LLC, St. Louis, Missouri 63105, United States; Department of Molecular Biosciences & Bioengineering, University of Hawaii at Manoa, Honolulu, Hawaii 96822, United States*; Email: groell@hawaii.edu

Yixin Chen − *Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*; Email: ychen25@wustl.edu

Yinjie J. Tang − *Department of Energy, Environmental, and Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*; orcid.org/0000-0002-5112-0649; Email: yinjie.tang@wustl.edu

Authors

Zhengyang Xiao − *Department of Energy, Environmental, and Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*; orcid.org/0000-0002-2617-2738

Wenyu Li − *Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*; orcid.org/0009-0009-0220-2259

Hannah Moon − *ImpactDB LLC, St. Louis, Missouri 63105, United States; Clayton High School, Clayton, Missouri 63105, United States*; orcid.org/0000-0001-8535-5246

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.3c00310

Author Contributions

Z.X. and W.L. contributed equally to this work. Conceptualization and Ideas − Y.J.T., Y.C., and G.W.R.; Methodology and Programming − GPT-4, W.L., and Z.X.; Data organization, curation, web site, and deposition − G.W.R., Z.X., and H.M.; Writing − ChatGPT, W.L., and Z.X.; Review & Editing − All authors.

Notes

The authors declare the following competing financial interest(s): G.W.R. has a financial interest in ImpactDB. The other authors declare no competing interests.

ACKNOWLEDGMENTS

This study is funded by United States NSF award number 2225809. The author GR was partially supported by the Defense Advanced Research Projects Agency B-SURE program (HR001122S0010). The views, opinions, and/or findings expressed should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

(1) Liao, X.; Ma, H.; Tang, Y. J. Artificial intelligence: a solution to involution of design-build-test-learn cycle. *Curr. Opin. Biotechnol.* **2022**, *75*, No. 102712.

(2) Czajka, J. J.; Oyetunde, T.; Tang, Y. J. Integrated knowledge mining, genome-scale modeling, and machine learning for predicting Yarrowia lipolytica bioproduction. *Metab. Eng.* **2021**, *67*, 227−236.

(3) Colletti, P. F.; Goyal, Y.; Varman, A. M.; Feng, X.; Wu, B.; Tang, Y. J. Evaluating factors that influence microbial synthesis yields by linear regression with numerical and ordinal variables. *Biotechnol. Bioeng.* **2011**, *108* (4), 893−901.

(4) Varman, A. M.; Xiao, Y.; Leonard, E.; Tang, Y. J. Statistics-based model for prediction of chemical biosynthesis yield from Saccharomyces cerevisiae. *Microb. Cell Fact.* **2011**, *10* (1), 45.

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

(6) Zhang, J.; Petersen, S. D.; Radivojevic, T.; Ramirez, A.; Pérez-Manríquez, A.; Abeliuk, E.; Sánchez, B. J.; Costello, Z.; Chen, Y.; Fero, M. J.; et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* **2020**, *11* (1), 4880.

(7) Roell, G. W.; Sathish, A.; Wan, N.; Cheng, Q.; Wen, Z.; Tang, Y. J.; Bao, F. S. A comparative evaluation of machine learning algorithms for predicting syngas fermentation outcomes. *Biochem. Eng. J.* **2022**, *186*, No. 108578.

(8) Radivojević, T.; Costello, Z.; Workman, K.; Garcia Martin, H. A machine learning Automated Recommendation Tool for synthetic biology. *Nat. Commun.* **2020**, *11* (1), 4879.

(9) Costello, Z.; Martin, H. G. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *npj Syst. Biol. Appl.* **2018**, *4* (1), 19.

(10) Winkler, J. D.; Halweg-Edwards, A. L.; Gill, R. T. The LASER database: Formalizing design rules for metabolic engineering. *Metab. Eng. Commun.* **2015**, *2*, 30−38.

(11) Wu, S. G.; Shimizu, K.; Tang, J. K. H.; Tang, Y. J. Facilitate collaborations among synthetic biology, metabolic engineering and machine learning. *ChemBioEng Rev.* **2016**, *3* (2), 45−54.

(12) Oyetunde, T.; Liu, D.; Martin, H. G.; Tang, Y. J.; Herrgard, M. Machine learning framework for assessment of microbial factory performance. *PLoS One* **2019**, *14* (1), No. e0210558.

(13) Zhu, J.-J.; Ren, Z. J. The evolution of research in resources, conservation & recycling revealed by Word2vec-enhanced data mining. *Resour. Conserv. Recycl.* **2023**, *190*, No. 106876.

(14) Jagannatha, A.; Liu, F.; Liu, W.; Yu, H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). *Drug Saf.* **2019**, *42* (1), 99−111.

(15) OpenAIGPT-4 Technical Report. *arXiv* 2023 DOI: 10.48550/arXiv.2303.08774

(16) Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv* 2023, DOI: 10.3390/ani13152515.

(17) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, 3, 9 DOI: 10.1186/s40537-016-0043-6.

(18) Nair, A. S.; Sivakumar, N. Enhanced production of biodiesel by Rhodosporidium toruloides using waste office paper hydrolysate as feedstock: Optimization and characterization. *Fuel* **2022**, *327*, No. 125174.

(19) Filippousi, R.; Diamantopoulou, P.; Stavropoulou, M.; Makris, D. P.; Papanikolaou, S. Lipid production by Rhodosporidium toruloides from biodiesel-derived glycerol in shake flasks and bioreactor: Impact of initial C/N molar ratio and added onion-peel extract. *Process Biochem.* **2022**, *123*, 52−62.

(20) Zheng, X.; Hu, R.; Chen, D.; Chen, J.; He, W.; Huang, L.; Lin, C.; Chen, H.; Chen, Y.; Zhu, J. Lipid and carotenoid production by the Rhodosporidium toruloides mutant in cane molasses. *Bioresour. Technol.* **2021**, *326*, No. 124816.

(21) Jiang, W.; Zhou, D.; Zhang, X.; Jiang, Y.; Zhang, W.; Xin, F.; Jiang, M. Co-production of lipids and carotenoids by Rhodosporidium toruloides from cane molasses using temperature and pH shifting strategies. *Biofuels Bioprod. Biorefin.* **2023**, *17*, 873.

(22) Wen, Z.; Zhang, S.; Odoh, C. K.; Jin, M.; Zhao, Z. K. Rhodosporidium toruloides - A potential red yeast chassis for lipids and beyond. *FEMS Yeast Res.* **2020**, *20* (5), No. foaa038. acccessed 3/9/2022

(23) Wang, K.; Johnson, C. W.; Bennett, K. C.; Johnson, P. A. Predicting fault slip via transfer learning. *Nat. Commun.* **2021**, *12* (1), 7319 DOI: 10.1038/s41467-021-27553-5.

(24) Cunha, A.; Pochet, A.; Lopes, H.; Gattass, M. Geo-sciences.Seismic fault detection in real data using transfer learning from a convolutional neural network pre-trained with synthetic seismic data. *Computers* **2020**, *135*, No. 104344.

(25) Rayner, K.; Schotter, E. R.; Masson, M. E.; Potter, M. C.; Treiman, R. So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help? *Psychol. Sci. Public Interest* **2016**, *17* (1), 4−34.

(26) Worland, A. M.; Czajka, J. J.; Li, Y.; Wang, Y.; Tang, Y. J.; Su, W. W. Biosynthesis of terpene compounds using the non-model yeast Yarrowia lipolytica: grand challenges and a few perspectives. *Curr. Opin. Biotechnol.* **2020**, *64*, 134−140.

(27) Ma, Y.; Liu, N.; Greisen, P.; Li, J.; Qiao, K.; Huang, S.; Stephanopoulos, G. Removal of lycopene substrate inhibition enables high carotenoid productivity in Yarrowia lipolytica. *Nat. Commun.* **2022**, *13* (1), 572 DOI: 10.1038/s41467-022-28277-w.

(28) Guo, Q.; Li, Y. W.; Yan, F.; Li, K.; Wang, Y. T.; Ye, C.; Shi, T. Q.; Huang, H. Dual cytoplasmic-peroxisomal engineering for high-yield production of sesquiterpene α-humulene in Yarrowia lipolytica. *Biotechnol. Bioeng.* **2022**, *119* (10), 2819−2830.

(29) Liu, G.-S.; Li, T.; Zhou, W.; Jiang, M.; Tao, X.-Y.; Liu, M.; Zhao, M.; Ren, Y.-H.; Gao, B.; Wang, F.-Q. The yeast peroxisome: a dynamic storage depot and subcellular factory for squalene overproduction. *Metab. Eng.* **2020**, *57*, 151−161.

(30) Ma, Y.; Li, J.; Huang, S.; Stephanopoulos, G. Targeting pathway expression to subcellular organelles improves astaxanthin synthesis in Yarrowia lipolytica. *Metab. Eng.* **2021**, *68*, 152−161.

(31) Chen, Z.; Zhao, P.; Li, F.; Marquez-Lago, T. T.; Leier, A.; Revote, J.; Zhu, Y.; Powell, D. R.; Akutsu, T.; Webb, G. I. J. B. i. biLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinformatics*20202*1*310471057 DOI: 10.1093/bib/bbz041

(32) Dong, G.; Duan, L.; Nummenmaa, J.; Zhang, P. Feature generation and feature engineering for sequences. In *Feature Engineering for Machine Learning and Data Analytics*; CRC Press, 2018; 145−166.

(33) Tran, T. N.; Ngo, D.-H.; Tran, Q. T.; Nguyen, H. C.; Su, C.-H.; Ngo, D.-N. Enhancing astaxanthin biosynthesis by Rhodosporidium toruloides mutants and optimization of medium compositions using response surface methodology. *Processes* **2020**, *8* (4), 497.

(34) Tran, T. N.; Quang-Vinh, T.; Huynh, H. T.; Hoang, N.-S.; Nguyen, H. C.; Dai-Nghiep, N. G. O. Astaxanthin production by newly isolated Rhodosporidium toruloides: optimization of medium compositions by response surface methodology. *Not. Bot. Horti Agrobot. Cluj-Napoca* **2019**, *47* (2), 320−327, DOI: 10.15835/nbha47111361.

(35) Tran, T. N.; Tran, N.-T.; Tran, T.-A.; Pham, D.-C.; Su, C.-H.; Nguyen, H. C.; Barrow, C. J.; Ngo, D.-N. Highly Active Astaxanthin Production from Waste Molasses by Mutated Rhodosporidium toruloides G17. *Fermentation* 2023, *9* (2), 148 DOI: 10.3390/fermentation9020148

(36) Niu, S.; Liu, Y.; Wang, J.; Song, H. A decade survey of transfer learning (2010−2020). *IEEE Trans. Artif. Intell.* **2020**, *1* (2), 151−166.

(37) He, K.; Zhang, X.; Ren, S.; Sun, J.Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv*2015; 1026−1034 DOI: 10.1109/ICCV.2015.123

(38) Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T. Palm-e: An embodied multimodal language model. *arXiv* 2023, DOI: 10.1002/anie.202302969.

(39) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Training language models to follow instructions with human feedback. *arXiv*, 2022, *35*, 27730−27744.

(40) Tramontin, L. R. R.; Kildegaard, K. R.; Sudarsan, S.; Borodina, I. Enhancement of astaxanthin biosynthesis in oleaginous yeast Yarrowia lipolytica via microalgal pathway. *Microorganisms* **2019**, *7* (10), 472.

(41) Zhu, H.-Z.; Jiang, S.; Wu, J.-J.; Zhou, X.-R.; Liu, P.-Y.; Huang, F.-H.; Wan, X. Production of High Levels of 3S,3′S-Astaxanthin in Yarrowia lipolytica via Iterative Metabolic Engineering. *J. Agric. Food Chem.* **2022**, *70* (8), 2673−2683.

(42) Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **2010**, *38* (suppl_1), D355−D360.