Generative artificial intelligence GPT-4 accelerates knowledge mining and machine learning for synthetic biology

Zhengyang Xiao[a,§], Wenyu Li[b,§], Hannah Moon[c,d], Garrett W. Roell[a,b,c,*], Yixin Chen[b,*], Yinjie J. Tang[a,*]

[a] Department of Energy, Environmental, and Chemical Engineering, Washington University in St. Louis, St. Louis, MO, 63130, USA

[b] Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, 63130, USA

[c] ImpactDB LLC. St. Louis, MO, 63105, USA

[d] Clayton High School, 1 Mark Twain Cir, Clayton, MO 63105, USA

[§]: These authors contributed equally to this work

***Corresponding Authors:**

Garrett Roell (garrettroell@wustl.edu)

Yixin Chen (ychen25@wustl.edu)

Yinjie Tang (yinjie.tang@wustl.edu)

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 greatly accelerate data extraction for machine learning to predict microbial performance under complex strain engineering and bioreactor conditions. As a proof of concept, GPT-4 was used to extract knowledge from 176 publications, resulting in 2037 data instances uploaded to a crowdsourcing online database. The centralized datasets and feature selection enabled a random forest model to predict fermentation titers of an industrial important yeast (*Yarrowia lipolytica*) with high accuracy (R^2 of 0.86 for unseen test data). Via transfer learning, the trained model could assess production capability of nonconventional yeasts (e.g., *Rhodosporidium toruloides*). This work showed the potential of generative AI to automate information extraction from research articles and advanced AI applications to facilitate design-build-test-learn (DBTL) for biomanufacturing as well as biotech commercial decisions.

Keywords: biomanufacturing, DBTL, feature selection, natural language processing, transfer learning, *Yarrowia lipolytica*

Introduction

Synthetic biology has historically relied on trial and error for breakthroughs due to the complex nature of biological systems. The development of artificial intelligence (AI) may improve DBTL cycles by learning from previously published results[1]. ML has been applied in synthetic biology to predict titers[2], optimize bioprocesses[3-5], and recommend strain engineering approaches[6, 7]. 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 supporting ML approaches. However, manually extracting data from a large number of articles is labor-intensive and prone to human errors[8], because reported data often lack a standardized format[9] and substantial efforts are needed to correctly interpret information and organize it into ML-ready data[10].

Fortunately, NLP, a branch of AI, can process text at a large scale, enabling topic organization in published articles[11]. It has also been utilized to track adverse drug events from electronic health record notes[12]. A recent tipping point in the field of NLP was the release of GPT-4[13], which shows 'sparks' of artificial general intelligence[14] to rapidly parse text based on user-provided context[13]. Leveraging GPT-4, relevant bioprocess features and outcomes (titer/rate/yield) from published papers can be extracted for rapid database growth. Since GPT-4 is unable to offer quantitative predictions of microbial productions (**Supplementary Fig. 1**), particularly for nonmodel species, this study has applied ML and transfer learning (TL) algorithms so that it is possible to predict novel microbial cell factories with limited data[2, 15].

As a proof of concept, this work demonstrates a workflow to extract data from articles on an industrial yeast *Yarrowia lipolytica* using GPT-4. The AI-extracted data can support ML methods to make accurate titer predictions based on numerical and categorical features of strain design, pathways, culture medium, and bioreactor conditions. Moreover, *Rhodosporidium toruloides* is a novel yeast that has recently gained research attention for its high lipid content[16, 17] and native carotenoid production[18, 19]. But the literature on *Rhodosporidium* is sparse[20]. Here, transfer learning (TL) can overcome this challenge by transferring knowledge from well-studied domains (*Yarrowia*-trained model) to understand less-studied scenarios, reducing computational costs and speeding up the learning process[21, 22]. Finally, all data extracted in this work was uploaded to an online database (impact-database.com) and the modeling programs are available on GitHub.

In summary, our study highlights the promising role of NLP in synthetic biology research, paving the way for more efficient data collection and rapid centralization of published case studies. The resulting database can support AI applications to assist biomanufacturing design, commercial decision, or project quality/risk assessment. This study, for the first time, demonstrates the integration of GPT with ML for predicting microbial cell factories. The lessons from this study will improve GPT applications via human supervision, prompt engineering, and reinforce learning.

Results and discussion

Extraction of ML features and datasets from synthetic biology papers.

Numerous Synthetic Biology tools have been developed to engineer microbes for sustainable

biomanufacturing. To develop microbial workhorses, iterative DBTL cycles have to be performed to overcome production bottlenecks, where model predictions of cell performance are keys to reduce experimental trials and improve strain development effectiveness. However, mechanistic models (such as genome scale model) have difficulties in incorporating all influential factors to simulate microbial productions under realistic bioreactor conditions[1], while ML is easier to be applied for modeling microbial cell factories by using the "knowledge" from previous lessons. The drawback of data-driven approaches is that it relies on a large amount of experimental data to correlate ML inputs (features) to outputs (productions). Since biomanufacturing literature presents a wealth of strain construction and bioprocess engineering case studies, the construction of a database from published papers may broadly support ML applications. On the other hand, data extraction without losing important knowledge is challenging because data reporting from publications is often sparse and inconsistent. Preferably, data mining should filter erroneous information and capture the features that independently affect bio-production. In this study, biomanufacturing features are grouped based on thermodynamics, reaction stoichiometry, bioreactor conditions, and genetic engineering methods (**Supplementary Table 1**). Manual extraction of these features and production outcomes can be time-consuming, and thus GPT-4 is used to overcome this challenge.

Data mining workflow using GPT-4.

A standardized workflow for GPT-4 data extraction and applications is shown in **Fig. 1**. The general principle is the division of complex tasks into small deliverables. Due to GPT-4's max

context window of 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 were then added to the beginning of each file depending on its contents (**Table 1**). We included some examples of GPT-4's response in **Supplementary File 1**. When the experiment results and methods are clearly conveyed in the text, GPT-4 summarizes the information into accessible tables. Although its output is not perfect, the results from GPT-4 allow users to verify data in a *targeted* manner, rather than laboriously reading each section[23]**.** While the AI output's accuracy is continuously improving, at this stage, human supervision is still a necessity. To extract data from one paper, the workflow typically took 20 minutes: labeling and dividing article sections→entering prompts on ChatGPT website→recording GPT response→combining the results into ML-ready dataset→quality check. This 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 easily checked for errors, and 4) it is adaptable to automation once the GPT-4 API is released.

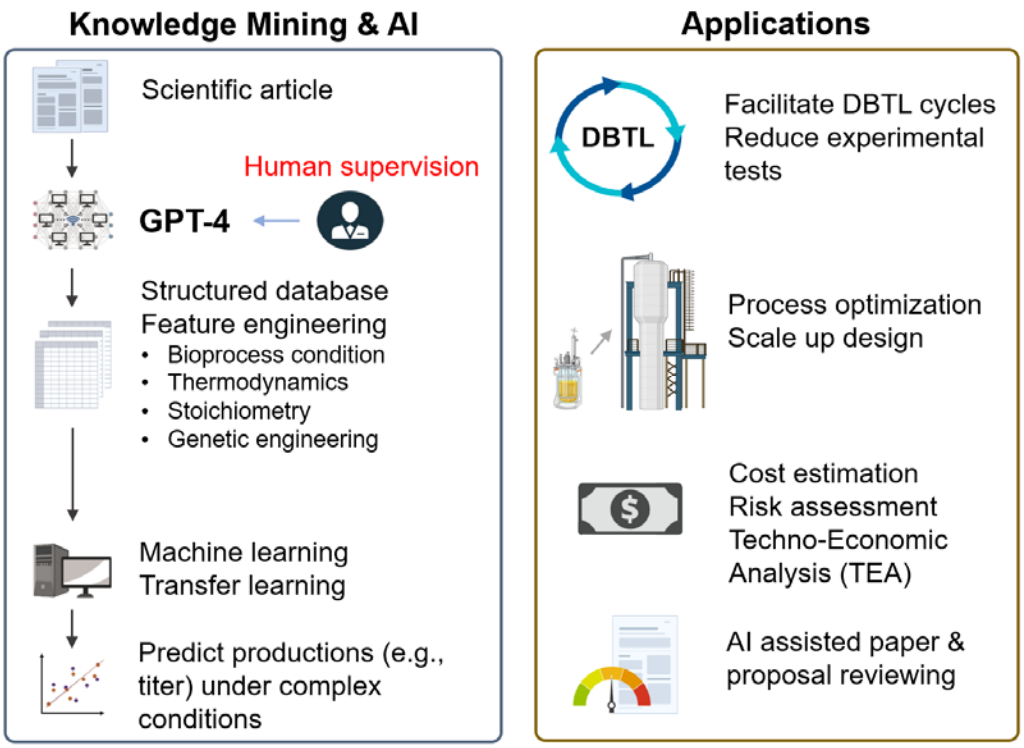

Fig. 1 GPT-4 knowledge mining and machine learning (ML).

Table 1 Prompts used for GPT-4 data extraction.

Prompt number	Information extracted	Section	Prompt text
1	Thermodynamics, growth condition	Abstract, Results	A multi-column spreadsheet from the given text | experiment number | strain name | product | product concentration | substrate | substrate concentration | media | time | mode of operation | + Abstract_text + Result_text
2	Genetic engineering	Materials and methods, Results	A multi-column 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

3	Growth condition	Materials and methods	A multi-column spreadsheet from the given text | experiment type | culture volume | culture vessel | substrate | substrate concentration | media | nitrogen condition | time | pH | temperature | + Method_text
4	Tabulated experiment data	Tables	Convert these into spreadsheets: + Table_text

Note: The redundant information in each prompt is necessary to match the bioprocess data in each instance and combine into the final ML-ready data table.

A case study for data extraction.

We started to use GPT-3.5 API on March 15th, 2023 to extract *Rhodosporidium* fermentation data from journal articles. The output of one PhD student using the GPT-enhanced workflow for one week is illustrated in Fig. 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 data extracted by GPT-3.5 in 10 papers. Compared with a previous manually extracted dataset[2], we found that titer data extracted by GPT-3.5 was 74% correct. Some of the erroneous data was obvious as they included numbers that were consecutive, repeated, or not present in the article. With minor user discretion to fix the outputs, the extracted titer data was found to be 89% correct. When GPT-4 extracted data from the test set of 10 *Yarrowia* papers, we found no errors in titer data. We summarized the data location, output format, and correctness in (**Supplemental Table 2**). In total, 366 data instances were obtained from 60 *Rhodosporidium* articles.

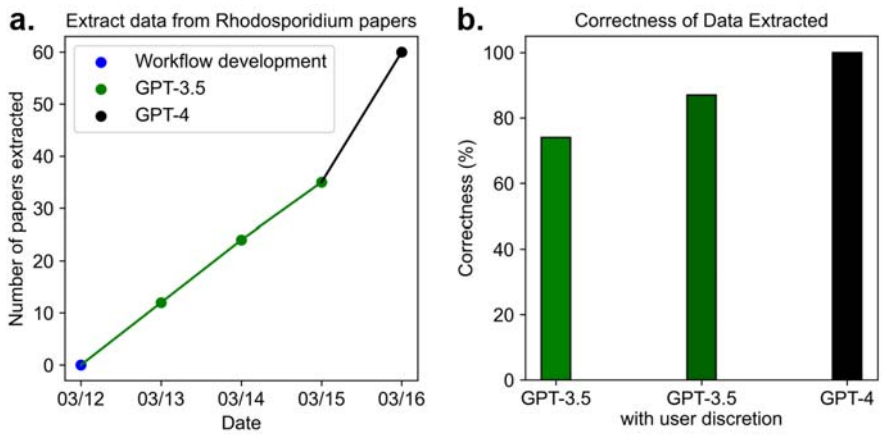

Fig. 2 Speed and correctness difference between GPT-3.5 and GPT-4. a. Number of *Rhodosporidium* papers processed in five days by a single user. **b.** The correctness of data extracted from a test set of 10 *Yarrowia* articles that contained 115 fermentation instances. Note: all data were manually inspected for error before use in machine learning.

GPT-4 assisted database construction for *Yarrowia lipolytica* biomanufacturing

Yarrowia lipolytica is an industrially important yeast for bio-productions[24]. Two years ago, it took over 400 work hours by Czajka et al. to extract information manually from ~100 *Yarrowia* papers (~3000 instances)[2]. Via GTP-4 workflow, ~1670 new data instances from 115 papers could be extracted within 40 working hours. We chose not to repeat the data extraction process for the articles in the Czajka et al. dataset since this would not produce new information. All extracted data for each fermentation instance was organized into 28 features that could be used for titer prediction (**Supplementary Table 1**). The manually extracted data was compared with GPT extracted data by computing feature importance, feature variances, and principal component analysis (PCA). The GPT extracted data had a similar distribution of feature importance to that of the manually extracted data (**Fig. 3a**), suggesting that the newly generated data followed similar patterns to the manually extracted data. Interestingly, the GPT datasets had higher feature variance than the manually extracted dataset for 19 of 28 features (**Fig. 3b**). Moreover, a

principal component analysis (PCA) showed that with K-means converged to the optimal solution and the same silhouette score, the data extracted by GPT showed 7% higher mean distance between clusters (**Fig. 4**). The PCA loadings further indicated that the clustering of the manually extracted dataset was governed mainly by carbon source and product cofactor cost (**Supplementary Fig. 2**). In contrast, GPT extracted data was clustered according to culture condition and genetic engineering features, in addition to carbon source and cofactor cost. These findings suggest that GPT-4 can capture more distinctiveness within papers and reason through complex textual data to generate less biased biomanufacturing instances.

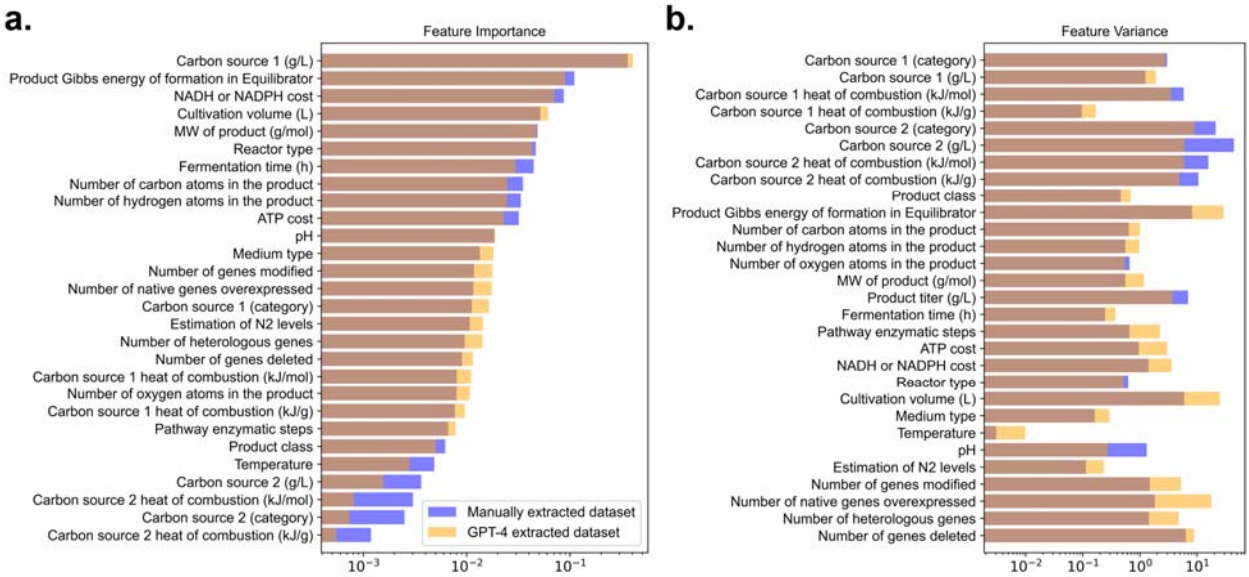

Fig. 3 Comparison of the manually extracted *Yarrowia* dataset with the GPT-4 extracted *Yarrowia* dataset. a. Feature importance as determined using a random forest regressor, ranked from high to low. **b.** Normalized feature variance.
Legend: Purple is the manually extracted dataset, and yellow is GPT-4 extracted dataset.

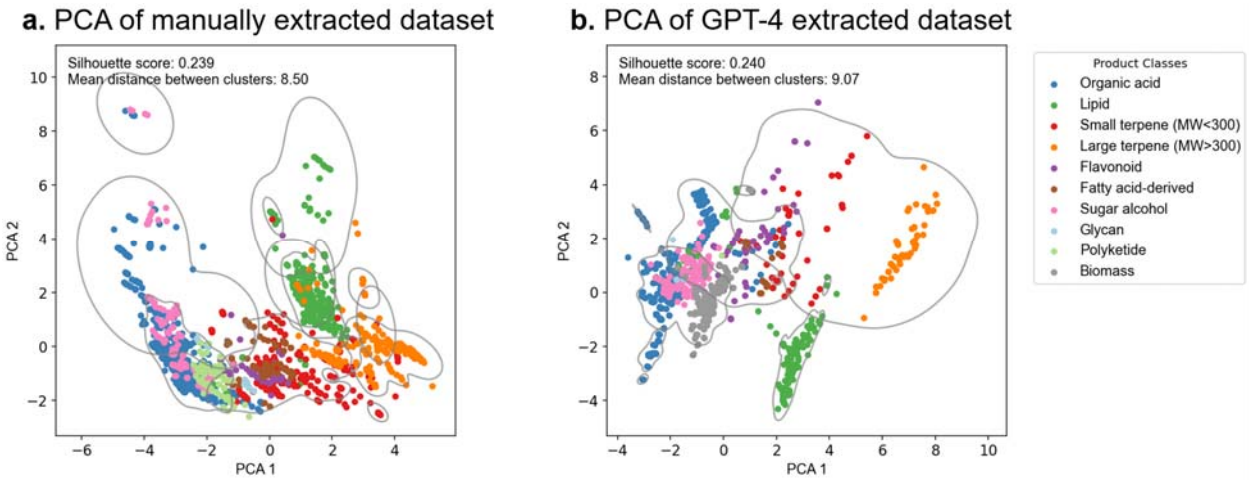

Fig. 4 PCA using K-means unsupervised learning. a) PCA of the manually extracted dataset. **b)** PCA of GPT-4 extracted dataset. Note the axis scale difference between panels **a** and **b**.

Leverage GPT database to predict *Y. lipolytica* fermentation titer.

Fermentation titer determines the bioprocess economy. Here, the GPT-constructed database can support the quantitative prediction of biomanufacturing. Specifically, the two *Y. lipolytica* datasets were combined to produce the most comprehensive *Y. lipolytica* bioproduction database to date. The data were used to conduct a comparative analysis of seven classical ML algorithms (**Supplementary Fig. 3**, support vector machine (SVM), gaussian process (GP), multi-layer perceptron (MLP), random forest (RF), extreme gradient boosting (XGBoost), k-nearest neighbors (KNN) and linear regression). Methods including linear regression and linear SVM performed the least well, suggesting that the titer prediction cannot be represented accurately by a linear relationship. A fully connected 2-layer neural network did not give the best performance either. Overall, the RF model achieved the best accuracy: R^2 of 0.86 on unseen test instances (**Fig. 5a**). Therefore, we chose to use the RF ensemble learner to predict production titer, based on a combination of substrate concentrations, thermodynamics data, growth conditions, and

engineered genes. After this point, we did not scale our train/test data, so their physical meaning was retained. The test set performance was robust, with an average R^2 of 0.80 across 50 random data splits. This represents a moderate improvement compared to Czajka et al., which reported an average R^2 of 0.76 in 10 random data splits ($p \leq 0.005$, student *t* test). The test set performance of the RF regressor was strong for nearly all product classes: organic acid, lipid, terpene, flavonoid, fatty acid-derived compound, sugar alcohol, glycan, and polyketide (**Fig. 5b-k**). The new *Y. lipolytica* ML model, trained on a database roughly 50% larger than the previous model, also showed general improvement for titer predictions of small terpene and polyketide products. However, the new model still struggles to explain the data related to large terpenes, as shown by the region with horizontal consecutive dots in the figure, where the same set of features results in a wide range of titers. This is because some key biomanufacturing features are still missing from knowledge mining. For example, the activities of key heterologous biosynthesis enzymes or bottleneck pathways were unable to be captured by the GPT.

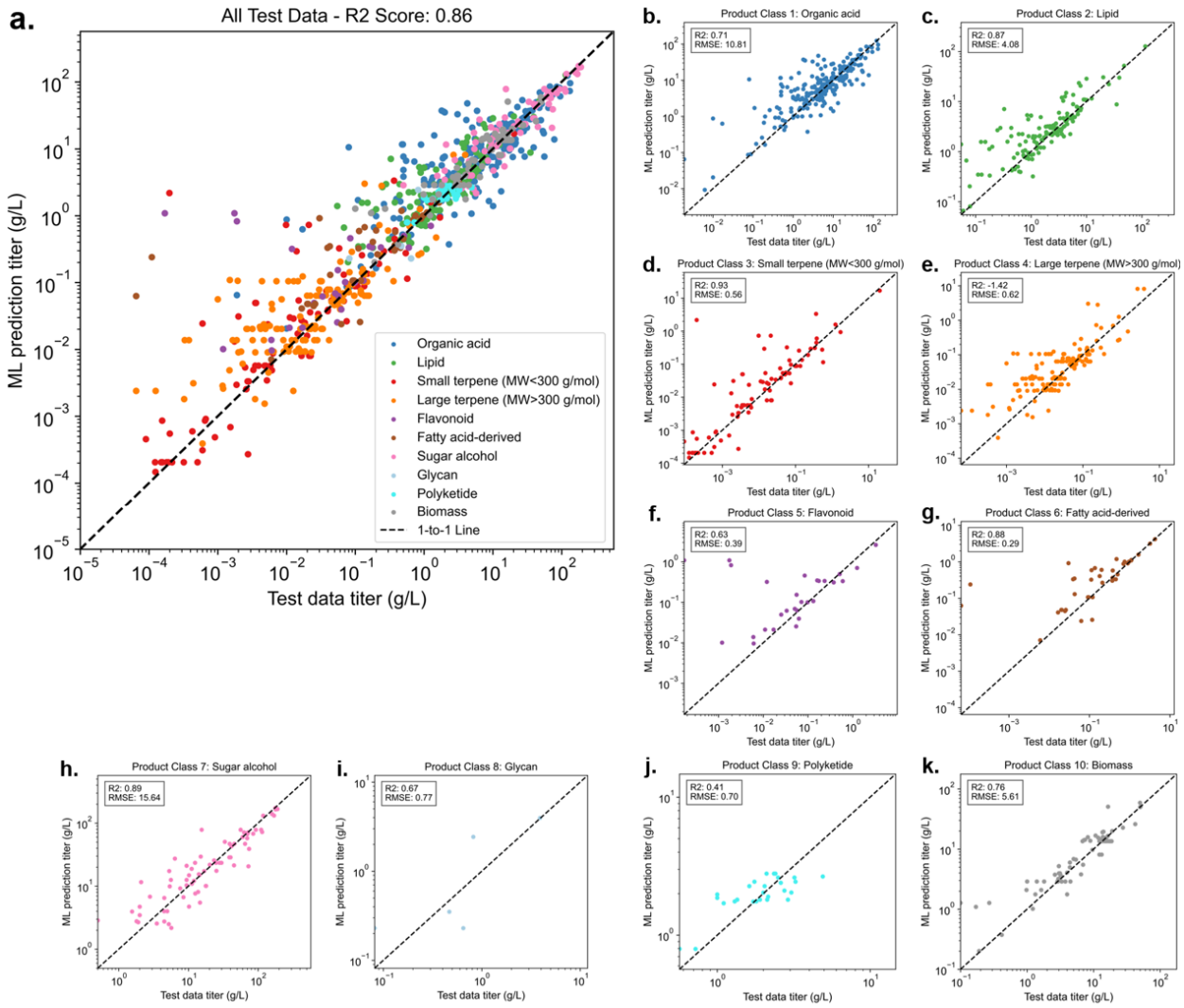

Fig 5. Test set *Y. lipolytica* titer predictions using a random forest ensemble learner.

Transfer learning from *Y. lipolytica* to *R. toruloides* for prediction of nonmodel yeast factories.

Non-model cell factories have been rapidly developing due to their unique metabolic capabilities. For example, *R. toruloides* is a nonmodel yeast that can convert cheap feedstock to high-value carotenoids. However, nonmodel cell factories have a limited number of reports. To better predict their performance under novel conditions, we utilized transfer learning to leverage

knowledge from the *Yarrowia* dataset. For example, our *Rhodosporidium* database contained 366 fermentation results extracted from 60 articles that could support RF model to predict lipid and biomass production (R^2 > 0.4) (**Supplementary Fig. 4**). But the database lacks genetic engineering features. For instance, reports on astaxanthin production in *R. toruloides* mainly focused on its native pathways[25-27]. Therefore, knowledge transferring from *Y. lipolytica* papers may predict how genetic engineering affects astaxanthin production in *R. toruloides* and offer guidelines for future strain development. Here, we utilized two inductive learning approaches[28]: 1) a neural network with an encoder-decoder structure to conduct presentation learning of the effect of the number of gene expressions on the astaxanthin synthesis; 2) an instance-based TL approach to address the source-target domain gap.

First, a pretrained encoder in an autoencoder[29] was used to minimize the effect of noise in the data through feature reduction. This method reduced the number of features from 29 (original 28 features + 1 categorical input of species, *Yarrowia or Rhodosporidium*) to 14. This model predicted the astaxanthin production after 96 hours of growth of *R. toruloides* in shake flasks containing YPD media. The predictions of the model trained on the encoded inputs were unsatisfactory because the model was insensitive to genetic modification features. Specifically, strains with 0, 2, or 4 heterologous gene overexpression were predicted to have similar astaxanthin titers. This indicates that encoding input data can impair the RF's ability to learn biologically significant knowledge among data, especially genetic modification features. Additionally, the stochastic nature of the encoder's neural network made this method insensitive to titer differences at the milligram level in a small dataset with little variance in features.

Therefore, we concluded that feature dimension reduction in TL may have a detrimental effect when applied to complex and interconnected biological systems.

As an alternative to the encoder approach, a RF transfer learning method was tested. The rationale for using a RF is that this algorithm has been shown to make accurate predictions from preliminary tests. During model training, the training instances were labeled as either *Yarrowia* or *Rhodosporidium* (categorical input feature), and a weight of 3x was assigned to the *Rhodosporidium* data. The model trained on data from both species was used to make *R. toruloides* astaxanthin titer predictions. Again, the inputs used corresponded to a 96-hour fermentation in shake flasks containing YPD media. The model predicted that wildtype *R. toruloides*, without heterologous genes, would produce an average astaxanthin titer of 4.2 mg/L, with most predictions below 2 mg/L (**Fig. 6**). This is comparable with a recent publication (not used in the model training) reporting astaxanthin production of 1.26 mg/L by *R. toruloides* in shake flasks[27]. Increasing the number of expressed genes from 0 to 6 resulted in increased titer predictions, but with diminishing returns for each additional heterologous gene. The highest titer predictions, an average titer of 39.5 mg/L, corresponded to over expressions of six heterologous genes. The broad distribution of astaxanthin production observed in the bootstrapping analysis suggests that genetic variations beyond the number of expressed genes play a crucial role in determining potential astaxanthin titer. For *Y. lipolytica*, high astaxanthin titer was typically achieved by employing heterologous enzymes with high activities[30, 31]. However, our current model is unable to account for molecular-level differences between enzymes of the same biosynthesis reaction. This RF transfer learning attempt is in essence a summary of strain

engineering experiences in *Y. lipolytica*, with transferred knowledge from the growth characteristics of *R. toruloides*.

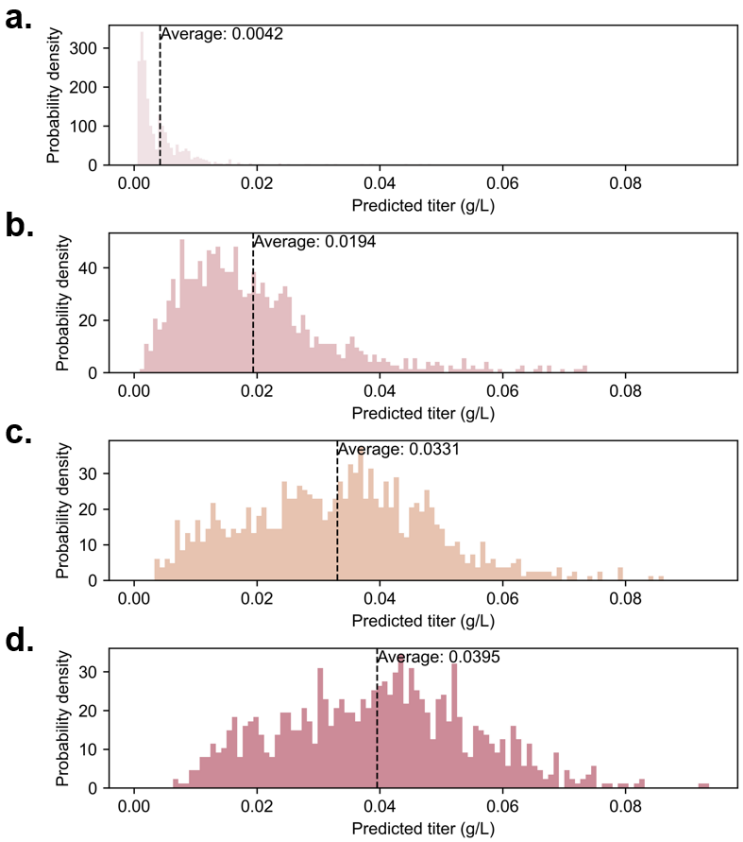

Fig 6. TL prediction for astaxanthin production in *R. toruloides*. a 0 heterologous gene expressed. **b** 2 heterologous genes expressed. **c** 4 heterologous genes expressed. **d** 6 heterologous genes expressed.

Biomanufacturing prediction quality.

For microbial hosts, such as *S. cerevisiae* and *E. coli*, holistic knowledge mining would require the almost impossible task of extracting discrete or continuous data from tens of thousands of relevant articles. Moreover, manual data extraction is prone to errors and inconsistencies in quality. The future releases of developer tools and plugins of GPT-4 have the potential to revolutionize data science in synthetic biology. AI can facilitate rapid data sensing, processing,

and the implementation of classical machine learning algorithms to accelerate the DBTL cycle. In this proof-of-concept study, the number of papers is only in the order of hundreds. If GPT-4 were to automate data collection, the collected data would approximate the entire population of data.

This study first demonstrated how GPT-4 can be used to extract data from synthetic biology publications. The data was organized into a format compatible with training and testing machine learning-based titer prediction models with 28 features. For *Yarrowia*, the majority of compounds could be predicted with relatively good accuracy (testing set $R^2 > 0.7$). However, the feature set used in this analysis is not complete and cannot adequately explain some outstanding works. 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[32]. However, the features in the collected database are not able to account for the substrate inhibition to production pathways. Also, our feature selection does not consider compartmentalization effects, as recent research showed that fermentation yields can be improved through cytoplasmic-peroxisomal engineering[33-35]. Additionally, the DNA sequence of key genes may be used as new ML inputs but still requires significant efforts for feature engineering[36, 37].

GPT-4's limitation and future direction.

Future large language models will have even more powerful data extraction abilities. For example, despite the token length limitation of around 8k, GPT-4 is able to capture contextual information effectively. Once higher token lengths of 32k become available, it will be possible to

input the entire article and fully automate data extraction. One current drawback is that graphical data is excluded from GPT-4's data extraction process. In the future, multimodal language models can interpret images for data extraction[13, 38]. Additionally, we experimented with GPT-4's ability to integrate data tables and output the final ML-ready dataset, and it showed satisfactory performance. However, due to token limitations, the process would stop after a few rows. So context length is a bottleneck for large-scale GPT applications. Besides, we observed other limitations in GPT-4 performance[39, 40]. According to OpenAI, GPT-4 has comparable biological science background knowledge to that of humans[13]. However, during our data extraction process, we encountered instances when GPT-4 was unable to differentiate between promoters and genes, or mixed up native and heterologous genes. Without knowledge about GPT-4's self-supervised learning mechanism, it is difficult to explain its performance because of its non-deterministic nature. Another common problem in large language models is that they provide seemingly plausible yet non-factual answers due to misunderstanding the prompt (i.e., questions for GPT). Although many attempts such as few-shot prompt engineering, Chain of Thought, and augmented language models have been proposed to address this issue[13], further case studies are still required to address the errors unique to synthetic biology. Therefore, we envision the future of automated data extraction to benefit greatly from reinforcement learning through human feedback, which fine tunes AI outputs and reduces errors[41].

ML algorithms and future applications.

GPT-4 can make decent predictions about common biomanufacturing products, such as lipids

(**Supplementary Fig. 1**). Nevertheless, it cannot make predictions beyond its database. Standalone TL is the solution when researching for the unknowns. As illustrated by our TL attempts, GPT-4 assisted programming enabled us to try both complicated and simple algorithms. With current features, the more complex and computationally expensive encoder-decoder structure had unsatisfactory results. In contrast, the simple RF with instance-transfer method exhibited poor precision but showed a good ability to give reasonable generalization predictions on unseen data. Most encoder-decoder structures for representation learning are designed to address larger databases. For tabular data in synthetic biology, the performance of neural networks is poor compared to other methods; thus, more specialized methods using deep learning can be developed to better represent tabular data. Moreover, we believe that incorporating more multi-omics features could improve the explainability of our model in the future[42-44], and AI plays a crucial role in acquiring sufficient training data. Our visions of future directions for AI-facilitated biomanufacturing include the prediction of cell production for techno-economic analyses (TEA) to inform commercialization decisions, as well as integration with computational strain design (CSD) for more effective strain engineering. Finally, we uploaded the AI-extracted data to a crowdsourcing online database ImpactDB so that researchers can easily contribute to the database and benefit the synthetic biology community. Ultimately, we envision that the development of NLP and AI will facilitate learning from past data and improve DBTL cycle.

Concluding remarks

Although artificial general intelligence is still in its infancy, it has shown great potential to

significantly assist and even revolutionize basic scientific research. Our work shows that it is feasible to leverage the power of large language models (LLMs) such as GPT-4 to automate knowledge mining from existing literature for supporting ML applications. AI models like GPT-4 can process vast amounts of information accurately, reducing the time and effort researchers need to spend on literature review and analysis. We also show that this ability, combined with transfer learning, can enable us to discover new knowledge on domains with limited data. We expect that generative AI has broad applications for biology research, including facilitating literature reviews, automating data analysis and interpretation, optimizing fermentations, predicting microbial outcomes, and ultimately accelerating the discovery and innovation in various fields.

Methods

GPT-3.5 and GPT-4 versions

The GPT-3.5 version[41] used was between the dates 3/10 to 3/15/2023. The GPT-4 version was between 3/16 to 4/2/2023.

Online database ImpactDB

The link to ImpactDB is https://impact-database.com/.

Data extraction from journal articles

Data extraction and feature organizations (**Supplementary Fig 1**) were done in a semi-automated fashion because currently GPT-4 is only available through OpenAI's ChatGPT website. First, text files for the different article sections (title, abstract, method, results, and text-based tables) were generated by manually combining the prompt sentence from **Table 1** with corresponding text that was copied-and-pasted from a manuscript. The prompt sentences were designed so that the extracted data from this study has the same format as the data from previous *Yarrowia* knowledge engineering effort[2]. GPT-4's responses were recorded and then later combined into a ML-ready format. The data correctness test was done by applying the above workflow to 10 articles of the *Yarrowia* paper database in ImpactDB. We excluded a few products not in the top ten product classes from our ML training due to their small sample volume. However, the data related to these products was uploaded to ImpactDB. The extracted data and paper number assignment are reported in **Supplementary File 2**. Standard condition thermodynamics data were obtained from NIST webbook (https://webbook.nist.gov/chemistry/). Non-standard thermodynamics data under biologically relevant concentrations were estimated by eQuilibrator 3.0 (https://equilibrator.weizmann.ac.il/[45]). Pathway enzyme steps, ATP, and NAD(P)H costs were estimated either by consulting KEGG[46] (https://www.genome.jp/kegg/) or reading the pathway map in relevant journal articles.

Manual and AI extracted data analysis, Data preprocessing, and ML

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 a RF regressor and extracting the corresponding weights. The clustering of PCA space is done by K-means unsupervised learning using a cluster number of 6, and Euclidean distances is 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. Categorical and numerical feature details are shown in **Supplemental File 2**. Stratified data split was performed to ensure that at least 20% data of each product class were included in the testing data. All data for ML models and TL are normalized using the min-max method to transform them between 0 and 1 in order to fit the TL normalization scale since one of the losses used is the MSE between input and embedding. Categorical data are encoded using ordinal encoding, enabling feature comparison after training and testing. Based on the correlation coefficients between features, for *R. toruloides*, carbon source concentration (g/L), carbon source's heat of combustion (kJ/L), and the NADH or NADPH cost have the strongest relationship with product titer, while for *Y. lipolytica* cultivation, volume and carbon source concentration (g/L) are strongly correlated with product titer. The parameters chosen for the preliminary ML regression task are shown in **Supplementary Table 3.**

TL via Pretrained-encoder and Random Forest with Instance-transfer

Following the same preprocessing procedure, we first train the autoencoder structure on all *Y. lipolytica* as source data, then transfer the encoder by freezing the layers to re-train it on *R.*

toruloides data without the Astaxanthin published data. The resulting embedding is then used to test a random forest model. The best result for predicting different numbers of heterologous genes by this method is shown in **Supplemental File 1** using the first model in **Table 2**. The resulting performance metrics for the regression task to predict all Astaxanthin product titer are shown in **Supplemental File 1**. The model architectures are summarized in the following table:

Table 2. Details of the encoder-decoder structure.

Model number	Loss equation	Model	Output dimension
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	2-layer encoder + 2-layer decoder with normalization	50% reduction
3	L_tot = L_Reconst + l1_weight *L_l1	2-layer encoder + 2-layer decoder with normalization	0% reduction

Note that all layers are fully connected, and early stopping is applied in both pretraining and retraining. Triplet loss is calculated as the Euclidean distances between the negative sample/positive sample and the anchor then applying the Rectified Linear Unit (ReLU) on the difference between the distances added by the margin. Reconstruction loss is calculated as the MSE between the original x and embedded x, and the l1 loss is the regression score calculated after encoding. Autoencoder weight is initialized using He initialization[29] for all experiments. The corresponding 0, 2, 4 heterologous gene expressions samples are encoded as input data and the prediction is performed by training a separate RF on a synthetic dataset obtained by bagging with replacement 100 times on all Astaxanthin data since the sample size for Astaxanthin is small.

The RF instance-transfer is done 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 dataset. The same data preprocessing was followed. The model is then tuned on the new training set and prediction is done using Astaxanthin samples with 0, 2, 4, 6 heterologous genes expressed respectively. The trained model was used to make predictions on the product titer. All codes are available at https://github.com/wenyuli23/SyntheticBiologyTL.

Statistical methods

All data were reported as the mean ± standard deviation. Statistical test used in this study was a one-tail student t-test.

Acknowledgements: This study is funded by United States NSF award number 2225809. The authors GR and YJT also received funding from Defense Advanced Research Projects Agency B-SURE program (HR001122S0010) for this research. 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.

Author contributions: Conceptualization and Ideas - YJT, YC and GR; Methodology and Programming – GPT-4, WL and ZX; Data organization, curation, website, and deposition - GR, ZX and HM; Writing – ChatGPT, WL and ZX; Review & Editing - YJT, YC, HM.

Supplementary files

Supplementary file 1: Supplementary figures and tables

Supplementary file 2: GPT-4-extracted database

References

(1) Liao, X.; Ma, H.; Tang, Y. J. Artificial intelligence: a solution to involution of design-build-test-learn cycle. *Current opinion in biotechnology* **2022**, *75*, 102712. DOI: 10.1016/j.copbio.2022.102712.

(2) Czajka, J. J.; Oyetunde, T.; Tang, Y. J. Integrated knowledge mining, genome-scale modeling, and machine learning for predicting Yarrowia lipolytica bioproduction. *Metabolic Engineering* **2021**, *67*, 227-236. DOI: https://doi.org/10.1016/j.ymben.2021.07.003.

(3) 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. *Nature Communications* **2022**, *13* (1), 541.

(4) 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. *Nature Communications* **2020**, *11* (1), 4880. DOI: 10.1038/s41467-020-17910-1.

(5) 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. *Biochemical Engineering Journal* **2022**, *186*, 108578. DOI: https://doi.org/10.1016/j.bej.2022.108578.

(6) Radivojević, T.; Costello, Z.; Workman, K.; Garcia Martin, H. A machine learning Automated Recommendation Tool for synthetic biology. *Nature Communications* **2020**, *11* (1), 4879. DOI: 10.1038/s41467-020-18008-4.

(7) Costello, Z.; Martin, H. G. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *npj Systems Biology and Applications* **2018**, *4* (1), 19. DOI: 10.1038/s41540-018-0054-3.

(8) Winkler, J. D.; Halweg-Edwards, A. L.; Gill, R. T. The LASER database: Formalizing design rules for metabolic engineering. *Metabolic Engineering Communications* **2015**, *2*, 30-38.

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

(10) Oyetunde, T.; Liu, D.; Martin, H. G.; Tang, Y. J. Machine learning framework for assessment of microbial factory performance. *PLoS One* **2019**, *14* (1), e0210558. DOI: 10.1371/journal.pone.0210558 From NLM.

(11) Zhu, J.-J.; Ren, Z. J. The evolution of research in resources, conservation & recycling revealed by Word2vec-enhanced data mining. *Resources, Conservation and Recycling* **2023**, *190*, 106876. DOI: https://doi.org/10.1016/j.resconrec.2023.106876.

(12) 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 Safety* **2019**, *42* (1), 99-111. DOI: 10.1007/s40264-018-0762-z.

(13) OpenAI. GPT-4 Technical Report. *arXiv preprint* **2023**, *2303.08774*.

(14) 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 preprint arXiv:2303.12712* **2023**.

(15) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *Journal of Big data* **2016**, *3* (1), 1-40.
(16) 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*, 125174.
(17) 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 Biochemistry* **2022**, *123*, 52-62.
(18) 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. *Bioresource Technology* **2021**, *326*, 124816.
(19) 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, Bioproducts and Biorefining* **2023**.
(20) 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 Research* **2020**, *20* (5), foaa038. DOI: 10.1093/femsyr/foaa038 (acccessed 3/9/2022).
(21) Wang, K.; Johnson, C. W.; Bennett, K. C.; Johnson, P. A. Predicting fault slip via transfer learning. *Nature Communications* **2021**, *12* (1), 7319.
(22) Cunha, A.; Pochet, A.; Lopes, H.; Gattass, M.; Geosciences. Seismic fault detection in real data using transfer learning from a convolutional neural network pre-trained with synthetic seismic data. *Computers* **2020**, *135*, 104344.
(23) 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? *Psychological science in the public interest : a journal of the American Psychological Society* **2016**, *17* (1), 4-34. DOI: 10.1177/1529100615623267 From NLM.
(24) 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. *Current opinion in biotechnology* **2020**, *64*, 134-140. DOI: https://doi.org/10.1016/j.copbio.2020.02.020.
(25) 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.
(26) 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. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* **2019**, *47* (2), 320-327.
(27) 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. **2023**, *9* (2), 148.
(28) Niu, S.; Liu, Y.; Wang, J.; Song, H. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence* **2020**, *1* (2), 151-166.
(29) He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. 2015, pp 1026-1034.

(30) 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.
(31) 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. *Journal of Agricultural and Food Chemistry* **2022**, *70* (8), 2673-2683. DOI: 10.1021/acs.jafc.1c08072.
(32) 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. *Nature Communications* **2022**, *13* (1), 572.
(33) 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. *Biotechnology and Bioengineering* **2022**, *119* (10), 2819-2830.
(34) 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. *Metabolic Engineering* **2020**, *57*, 151-161.
(35) Ma, Y.; Li, J.; Huang, S.; Stephanopoulos, G. Targeting pathway expression to subcellular organelles improves astaxanthin synthesis in Yarrowia lipolytica. *Metabolic Engineering* **2021**, *68*, 152-161.
(36) 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. b. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. **2020**, *21* (3), 1047-1057.
(37) 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; pp 145-166.
(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 preprint arXiv:2303.03378* **2023**.
(39) Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. **2023**.
(40) Lubiana, T.; Lopes, R.; Medeiros, P.; Silva, J. C.; Goncalves, A. N. A.; Maracaja-Coutinho, V.; Nakaya, H. I. Ten Quick Tips for Harnessing the Power of ChatGPT/GPT-4 in Computational Biology. *arXiv preprint arXiv:2303.16429* **2023**.
(41) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A. J. A. i. N. I. P. S. Training language models to follow instructions with human feedback. **2022**, *35*, 27730-27744.
(42) Reel, P. S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances* **2021**, *49*, 107739.
(43) Lin, E.; Lane, H.-Y. Machine learning and systems genomics approaches for multi-omics data. *Biomarker research* **2017**, *5*, 1-6.
(44) Picard, M.; Scott-Boyer, M.-P.; Bodein, A.; Périn, O.; Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal* **2021**, *19*, 3735-3746.

(45) Beber, M. E.; Gollub, M. G.; Mozaffari, D.; Shebek, K. M.; Flamholz, A. I.; Milo, R.; Noor, E. eQuilibrator 3.0: a database solution for thermodynamic constant estimation. *Nucleic acids research* **2022**, *50* (D1), D603-D609.
(46) Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research* **2010**, *38* (suppl_1), D355-D360.