Comparative Evaluation of Machine Learning Models for Predicting Biomass Production in Thraustochytrid Cultivation

Farzane Nourmand[1], * Elham Iranmanesh[2], Masoud Torkzadeh-Mahani[1], Esmat Rashedi[3], Shahryar Shakeri[1]
[1] Department of Biotechnology, Institute of Science, High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran, farzane.nourmand@gmail.com, mtmahani@gmail.com, shahryar.shakeri@yahoo.com
[2] Department of Chemical Engineering, Faculty of Chemistry and Chemical Engineering, Graduate University of Advanced Technology, Kerman, Iran, e.iranmanesh@kgut.ac.ir
[3] Department of Communication and Electrical Engineering, Faculty of Electrical and Computer Engineering, Graduate University of Advanced Technology, Kerman, Iran, esmat.rashedi@gmail.com

ABSTRACT

Thraustochytrids are promising marine microorganisms for the production of high-value compounds such as docosahexaenoic acid (DHA). However, optimizing their cultivation conditions using conventional experimental approaches is labor-intensive and time-consuming. In this study, the performance of Support Vector Regression and Random Forest models was compared for predicting biomass production under different cultivation conditions. Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Among the evaluated models, SVR demonstrated superior predictive performance, achieving higher prediction accuracy and lower error values than RF. These findings suggest that SVR is a reliable tool for biomass prediction and can facilitate the optimization of Thraustochytrid cultivation while reducing experimental effort and cost.

Keywords: Thraustochytrids, Machine Learning, Support Vector Regression, Random Forest, Biomass Prediction.

1. INTRODUCTION

Thraustochytrids belong to a group of unicellular, heterotrophic marine protists that represent the phylum Stramenopila, class Labyrinthulomycetes, and family Thraustochytriaceae [1], [2]. Historically, they were classified as algae due to their saprobic lifestyle; however, gene tree topology based on 18S rRNA phylogenetic analysis revealed their striking difference on the genetic level [3]. Regarding their morphology, *Thraustochytrids* possess a complex life history that includes a polymorphic stage of actively motile biflagellate zoospores alternating with the coccoid form of the vegetative cells[1]. At ultrastructural levels, *Thraustochytrids* have a complicated anatomy that comprises lipid droplets, a bothrosome, and a sophisticated ectoplasmic network of vesicular appendages involved in the secretion of digestive enzymes and the degradation of dissolved organic matter deriving from mangrove leaves and estuarine sediments[3].

Thraustochytrids have emerged as promising commercial cell factories from a biotechnological perspective[4]. Their interest in industry is mainly due to their capacity to produce high amounts of valuable neutral lipids, mostly triacylglycerols, which can account for up to 60% of the total dry weight of their cells when induced by specific stressful conditions[5], [6]. Importantly, up to 50% of this intracellular lipid pool consists of docosahexaenoic acid (DHA, C22:6 ω-3) and eicosapentaenoic acid (EPA, C20:5 ω-3), which are essential long-chain polyunsaturated fatty acids important for cardiovascular health, neural development in

infants, and the prevention of various inflammatory diseases[4], [6]. *Thraustochytrids* also possess a highly functional secondary metabolism, in which they accumulate appreciable levels of the natural triterpenoid squalene, a compound of high value in pharmaceuticals and cosmetics, and a range of antioxidant carotenoid pigments such as astaxanthin, β-carotene, canthaxanthin, and phoenicoxanthin[1], [3]. They are a toxin-free and sustainable alternative to fish oil, growing quickly and easily cultivated in artificial environments[6], [7]. Despite their commercial promise, the scalability of *Thraustochytrid* fermentation is constrained by the complex, multivariable interactions governing biomass and metabolite accumulation[8]. The cultivation process is characterized by complex, non-linear metabolic shifts influenced by environmental variables such as dissolved oxygen, pH, temperature, and substrate concentration[5].

Traditionally, bioprocess optimization has relied on static response surfaces or mechanistic kinetic models such as the substrate-independent logistic and Luedeking-Piret equations[3]. However, these classical techniques often fail to account for the extreme non-linear metabolic shifts and local optima that are characteristic of dynamic microbial ecosystems[3], [5]. In addition, mechanistic and genome-scale metabolic models (GEMs) need the calibration of many parameters and do not have the capacity to adapt in real-time to variations at the industrial scale[8]. This optimization gap motivates a shift from rigid mechanistic frameworks to flexible data-driven modeling approaches[3]. Machine learning (ML) provides a transformative solution to this challenge by enabling algorithms to improve performance through the identification of patterns directly from cultivation data, without the need for explicit mechanistic programming[9], [10]. In bioprocessing, mach ML architectures circumvent the rigidity of traditional kinetic models by capturing the intricate, multi-variable interactions governing growth and metabolite titers[11]. While diverse algorithms are available, this study focuses specifically on two robust paradigms particularly suited for bioprocessing data: Random Forest (RF) and Support Vector Regression (SVR)[12].

RF is a bagging-based ensemble method that constructs an array of uncorrelated decision trees, averaging their predictions to reduce model variance and prevent overfitting, a critical advantage when dealing with the limited data density typical of laboratory-scale research[12]. Parallel to this, SVR utilizes hyperplane classifiers and kernel functions, such as the Radial Basis Function (RBF), to map input features into a multidimensional space where a linear hyperplane can be used to resolve complex, nonlinear metabolic patterns[12]. The primary objective of this study is to conduct a comparative evaluation exclusively between these two specific architectures, RF and SVR, to identify the most effective tool for predicting and optimizing biomass production in *Thraustochytrid* cultivation[13]. By benchmarking the predictive accuracy and robustness of only RF and SVR using key statistical indicators, including the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), this research establishes a data-driven framework that enhances bioprocess efficiency[12], [13]. Identifying the superior algorithm between these two is a vital step toward the digital transformation of *Thraustochytrid* bioprocessing, ensuring more sustainable and economically viable production of high-value marine lipids[3].

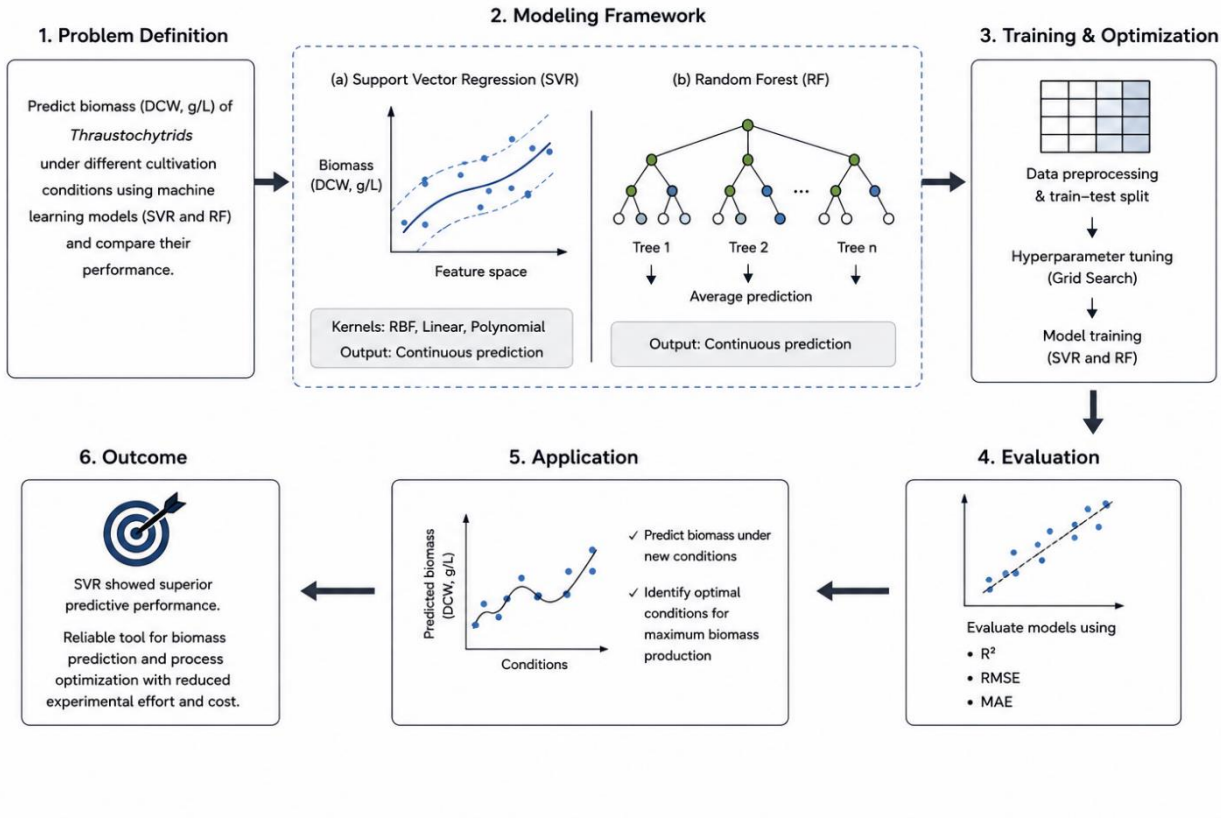

Fig.1. *Workflow of the proposed ML framework*

2. METHODS

2.1 Data Set and Feature Selection

The empirical basis of the study is 39 sets of experimental observations, systematically obtained from published research on *Thraustochytrid* cultivation. The ML models are developed based on three main input features, namely, carbon to nitrogen (C/N) ratio, salinity, and impeller agitation speed (RPM)[3], [9], [21], [22](Fig.1). These bioprocess parameters were selected because of their well-defined non-linear impact on the cellular physiology and metabolic changes of *Thraustochytrids*. All predictive algorithms are trained to predict final biomass production in g/LDCW (grams per liter of dry cell weight)[13], [19], [20].

2.2 Data Preprocessing and Model Partitioning

Z-score normalization was conducted on the raw data set to standardize the variables with different numerical scales to ensure the stability and accuracy of the model[12]. To evaluate the predictive performance of the models and their ability to generalize to unseen data, the dataset was divided with a 70-30 train/test split. Specifically, 70% of the data was utilized for training and hyperparameter tuning, and the other 30% was set aside as an independent test set for validation[12].

2.3 Model Implementation and Performance Benchmarks

The complex relation between cultivation features and biomass yield is modeled using two different ML paradigms, namely RF and SVR. RF is an ensemble bagging method to reduce the variance and prevent overfitting. SVR uses kernel functions to map data to a high-dimensional space for complex regression[11].

3. RESULTS AND DISCUSSION

3.1 Statistical Performance Evaluation

The predictive performance of the RF and SVR models was evaluated using the R^2, RMSE, and MAE across the 70-30 train/test split. The results are summarized in Table 1

Table 1. *Comparison of the performance of SVR and RF*

model	Training R^2	Testing R^2	RMSE	MAE
SVR	0.900	0.916	0.069	0.067
RF	0.830	0.663	0.139	0.112

The SVR model demonstrated the best predictive accuracy and robustness among the other models, with an R^2 value of 0.9168 for the independent test set. The RF model, on the other hand, had a large difference between training (0.8307) and testing (0.6683) accuracy, as well as larger errors (RMSE: 0.1396 vs. 0.0699). This suggests that for this particular *Thraustochytrid* data set, the kernel-based approach of SVR achieved more accurate biomass estimates with about half the absolute error of the tree-based ensemble method.

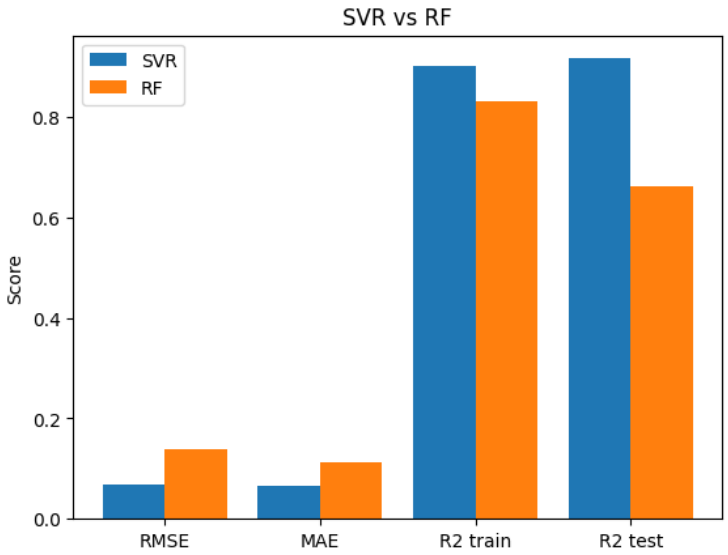

Fig. 2. *Comparison of the predictive performance of SVR and RF models based on R², RMSE, and MAE*

The graphical comparison in Fig. 2 further reveals the superiority of the SVR model over RF. The SVR model was found to be more accurate and robust in predicting the experimental dataset with a higher coefficient of determination and significantly lower prediction errors.

3.2 Comparative Discussion

The comparison of both algorithms shows that SVR provided a more reliable prediction model for biomass estimation under the investigated cultivation conditions. The higher coefficient of determination, lower RMSE and MAE values indicated that SVR was more successful in modeling the nonlinear relation between cultivation variables and biomass production. The excellent performance of SVR can be explained by its

powerful nonlinear regression ability and structural risk minimization principle to enhance the generalization ability of the model[9]. These features make SVR particularly appropriate for biological datasets, where nonlinear interactions between environmental and nutritional factors are common. RF is known to be a strong ensemble learning algorithm, which can model complex relationships[12]. However, in this study, RF was found to be worse in prediction than SVR. The lower testing R^2 and higher prediction errors indicated a lower success of RF in accurately estimating biomass production across different cultivation conditions.

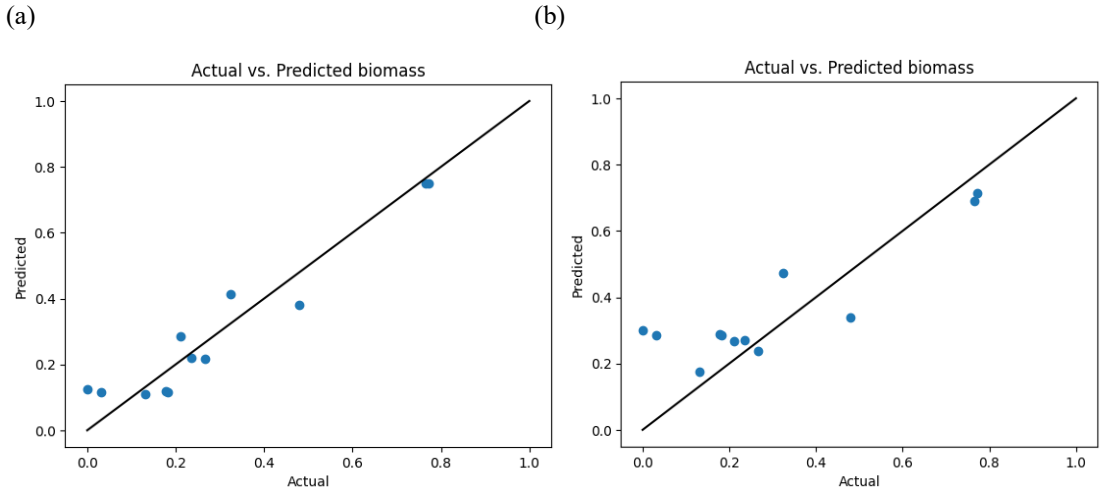

Fig. 3. *Comparison of predicted and actual values in (a) SVR and (b) RF algorithms.*

The scatter plot in Fig.3 shows a close agreement between predicted and experimental biomass values. Most data have high prediction accuracy, and most data are close to the diagonal reference line, indicating good reliability of the model. The RF predictions diverge more from the reference line than the SVR model, indicating a lower predictive accuracy and larger estimation errors. Data-driven optimization of *Thraustochytrid* cultivation has advanced significantly, as evidenced by the predictive performance of the models created in this study, especially the SVR. With a testing R^2 of 0.916, our SVR model significantly outperforms conventional statistical optimization methods.

For larger datasets, like the cultivation of *Chlorella vulgaris* (N=145), RF was the best predictor with an R^2 of 0.907[14]. Similarly, in nutrient requirement modeling for *Dunaliella*, ANN performed better than SVR with increasing data size[15]. However, for the less extensive experimental dataset used in this study (N=43), our results agree with the previous findings that tree-based methods are unable to capture the noise rather than biological signals when data is sparse and thus overfit, as demonstrated by the performance loss of 16.7% of our RF model from training to testing phases. Moreover, our results provide a competitive benchmark against more complex deep learning architectures that require high-frequency data. Although advanced models such as ANN-GA or LSTM-Attention have achieved extremely high predictive accuracies ($R^2 > 0.98$) for *Schizochytrium* sp., those models are based on multi-stage dynamic fermentation data or multi-sensor time-series that are often not available in early-stage laboratory R&D[15], [16]. Our SVR approach, in contrast, provides a robust alternative for typical laboratory-scale datasets where N<100. Our reported RMSE of 0.069 is also consistent with the high precision observed in complex “soft sensor” frameworks for *Schizochytrium* sp[17]. with range-normalized errors as low as 1.1%. These results indicate that the SVR model is a good basis for real-time monitoring and digital transformation of bioprocessing[15].

3.3 Practical Implications

The results showed that ML algorithms can predict the biomass production in *Thraustochytrid* cultivation well, and the need for extensive laboratory experiments can be reduced. SVR yielded the highest prediction

accuracy and the lowest prediction errors among the tested models and was found to be a useful computational tool for rapid bioprocess optimization. The availability of accurate predictive models can minimize the costs related to experimentation and speed up the optimization of cultivation conditions for industrial biomass production[3], [18].

CONCLUSION

This study shows that SVR provides a very robust and accurate computational framework for the prediction of biomass production in *Thraustochytrid* cultivation. The comparative evaluation shows that the kernel-based SVR model performed much better than the RF ensemble approach, with an excellent testing R^2 of 0.916 and a low RMSE of 0.069. The high accuracy of SVR prediction suggests that C/N ratio, salinity, and agitation speed are the dominant factors driving the growth of *Thraustochytrids* and can be used to model the complex non-linear metabolic transitions in these organisms. These results offer a strong data-driven foundation for the development of soft sensors and digital twins in the field, allowing real-time monitoring of active biomass titers by means of easily measurable online variables. Future work should focus on the expansion of this dataset and the development of hybrid models combining the mechanistic kinetic equations with the adaptive capabilities of the neural networks to improve the scalability and sustainability of high-value marine lipid production.

REFERENCES

[1] *A. Leyton et al., "Antarctic Thraustochytrids as Sources of Carotenoids and High-Value Fatty Acids," Marine Drugs, vol. 19, no. 7, p. 386, Jul. 2021, doi: 10.3390/md19070386.*

[2] *Q. Wang, Y. Zhang, R. Hui, and Y. Zhu, "Marine thraustochytrid: exploration from taxonomic challenges to biotechnological applications," Front. Mar. Sci., vol. 11, p. 1371713, Mar. 2024, doi: 10.3389/fmars.2024.1371713.*

[3] *A. G. Menzorov et al., "Thraustochytrids: Evolution, Ultrastructure, Biotechnology, and Modeling," IJMS, vol. 25, no. 23, p. 13172, Dec. 2024, doi: 10.3390/ijms252313172.*

[4] *Y. Zhao, X. Zhu, N. Riaz, X. Liu, J. Li, and G. Wang, "Salinity Modulates Carbon Flux to Promote Squalene and PUFA Biosynthesis in the Marine Protist Thraustochytrium," Marine Drugs, vol. 23, no. 9, p. 354, Aug. 2025, doi: 10.3390/md23090354.*

[5] *Z. Bartosova, H. Ertesvåg, E. L. Nyfløt, K. Kämpe, I. M. Aasen, and P. Bruheim, "Combined Metabolome and Lipidome Analyses for In-Depth Characterization of Lipid Accumulation in the DHA Producing Aurantiochytrium sp. T66," Metabolites, vol. 11, no. 3, p. 135, Feb. 2021, doi: 10.3390/metabo11030135.*

[6] *H. Zhong, W. Liu, and Y. Tao, "Genomic and Transcriptomic Characterization of a High-Yield Docosahexaenoic Acid (DHA) Mutant Schizochytrium sp. HS01," Fermentation, vol. 11, no. 11, p. 631, Nov. 2025, doi: 10.3390/fermentation11110631.*

[7] *L. Liang et al., "Genome and Transcriptome Analyses Provide Insight Into the Omega-3 Long-Chain Polyunsaturated Fatty Acids Biosynthesis of Schizochytrium limacinum SR21," Front. Microbiol., vol. 11, p. 687, Apr. 2020, doi: 10.3389/fmicb.2020.00687.*

[8] *V. Simensen, A. Voigt, and E. Almaas, "High-quality genome-scale metabolic model of Aurantiochytrium sp. T66," Biotech & Bioengineering, vol. 118, no. 5, pp. 2105–2117, May 2021, doi: 10.1002/bit.27726.*

[9] *Z.-Z. Wang et al., "Fermentation design and process optimization strategy based on machine learning," BioDesign Research, vol. 7, no. 1, p. 100002, Mar. 2025, doi: 10.1016/j.bidere.2025.100002.*

[10] *Z. Pan, "Machine Learning for Real-time Optimization of Bioprocessing Parameters: Applications and Improvements".*

[11] *X.-Y. Huang et al., "Developing high-dimensional machine learning models to improve generalization ability and overcome data insufficiency for mixed sugar fermentation simulation," Bioresource Technology, vol. 385, p. 129375, Oct. 2023, doi: 10.1016/j.biortech.2023.129375.*

[12] *D. Campos, R. Ervilha, M. Bodini, C. M. Saporetti, and L. Goliatt, "Hybrid evolutionary machine learning framework optimizing biochar production in biomass pyrolysis," Fuel Processing Technology, vol. 284, p. 108419, Jun. 2026, doi: 10.1016/j.fuproc.2026.108419.*

[13] *Acosta-Pavas Juan C. et al., "Soft-sensors Strategy for Estimating the Cell Biomass and the Nanobody-based Scorpion Antivenom Production: from Ml Algorithms Selection to Scale-down Bioprocess Application," Chemical Engineering Transactions, vol. 119, pp. 349–354, Nov. 2025, doi: 10.3303/CET25119059.*

[14] *A. P. Peter et al., "Artificial intelligence model for monitoring biomass growth in semi-batch Chlorella vulgaris cultivation," Fuel, vol. 333, p. 126438, Feb. 2023, doi: 10.1016/j.fuel.2022.126438.*

[15] *E. Imamoglu, "Artificial Intelligence and/or Machine Learning Algorithms in Microalgae Bioprocesses," Bioengineering, vol. 11, no. 11, p. 1143, Nov. 2024, doi: 10.3390/bioengineering11111143.*

[16] *R. Ervilha et al., "Bayesian optimization-guided machine learning models for bio-oil yield prediction from lignocellulosic biomass pyrolysis," Biomass and Bioenergy, vol. 203, p. 108280, Dec. 2025, doi: 10.1016/j.biombioe.2025.108280.*

[17] *C. Alarcon and C. Shene, "Arduino Soft Sensor for Monitoring Schizochytrium sp. Fermentation, a Proof of Concept for the Industrial Application of Genome-Scale Metabolic Models in the Context of Pharma 4.0," Processes, vol. 10, no. 11, p. 2226, Oct. 2022, doi: 10.3390/pr10112226.*

[18] *U. C. Nneoma, O. F. Chukwudi, U. J. Nnenna, and U. O. Paul-Chima, "Hybrid biofactories: integrating microalgae and engineered microbiomes for enhanced biofuel production in circular carbon systems," Front. Energy Res., vol. 13, p. 1654079, Sep. 2025, doi: 10.3389/fenrg.2025.1654079.*

[19] *J. Zhao, D. Chen, C. Lin, and X. Wan, "Enhancing astaxanthin production in Schizochytrium sp.: Insights from orthogonal experiments and transcriptome analysis," Algal Research, vol. 86, p. 103966, Mar. 2025, doi: 10.1016/j.algal.2025.103966.*

[20] *R. Xiao, X. Li, E. Leonard, N. Tharayil, and Y. Zheng, "Investigation on the effects of cultivation conditions, fed-batch operation, and enzymatic hydrolysate of corn stover on the astaxanthin production by Thraustochytrium striatum," Algal Research, vol. 39, p. 101475, May 2019, doi: 10.1016/j.algal.2019.101475.*

[21] *Q. Zhang, D. Deng, W. Dai, J. Li, and X. Jin, "Optimization of culture conditions for differentiation of melon based on artificial neural network and genetic algorithm," Sci Rep, vol. 10, no. 1, p. 3524, Feb. 2020, doi: 10.1038/s41598-020-60278-x.*

[22] *K. Khumrangsee, T. Charoenrat, J. Praiboon, and S. Chittapun, "Development of a fed-batch fermentation with stepwise aeration to enhance docosahexaenoic acid and carotenoid content in Aurantiochytrium sp. FIKU018," J Appl Phycol, vol. 34, no. 3, pp. 1243–1253, Jun. 2022, doi: 10.1007/s10811-022-02726-x.*