

Article

Artificial Neural Network Elucidates the Role of Transport Proteins in *Rhodopseudomonas palustris* CGA009 During Lignin Breakdown Product Catabolism

Niaz Bahar Chowdhury, Mark Kathol, Nabia Shahreen  and Rajib Saha * 

Chemical and Biomolecular Engineering, University of Nebraska-Lincoln, Lincoln, NE 68508, USA; nchowdhury2@huskers.unl.edu (N.B.C.); nshahreen2@huskers.unl.edu (N.S.)

* Correspondence: rsaha2@unl.edu

Abstract

Background: *Rhodopseudomonas palustris* is a metabolically versatile bacterium with significant biotechnological potential, including the ability to catabolize lignin and its heterogeneous breakdown products. Understanding the molecular determinants of growth on lignin-derived compounds is essential for advancing lignin valorization strategies under both aerobic and anaerobic conditions. **Methods:** *R. palustris* was cultivated on multiple lignin breakdown products (LBPs), including *p*-coumaryl alcohol, coniferyl alcohol, sinapyl alcohol, *p*-coumarate, sodium ferulate, and kraft lignin. Condition-specific transcriptomics and proteomics datasets were generated and used as input features to train machine-learning models, with experimentally measured growth rates as the prediction target. Artificial Neural Networks (ANNs), Random Forest (RF), and Support Vector Machine (SVM) models were evaluated and compared. Permutation feature importance analysis was applied to identify genes and proteins most influential for growth. **Results:** Among the tested models, ANNs achieved the highest predictive performance, with accuracies of 94% for transcriptomics-based models and 96% for proteomics-based models. Feature importance analysis identified the top twenty growth-associated genes and proteins for each omics layer. Integrating transcriptomic and proteomic results revealed eight key transport proteins that consistently influenced growth across LBP conditions. Re-training ANN models using only these eight transport proteins maintained high predictive accuracy, achieving 86% for proteomics and 76% for transcriptomics. **Conclusions:** This study demonstrates the effectiveness of ANN-based models for predicting growth-associated genes and proteins in *R. palustris*. The identification of a small set of key transport proteins provides mechanistic insight into lignin catabolism and highlights promising targets for metabolic engineering aimed at improving lignin utilization.

Keywords: *Rhodopseudomonas palustris*; lignin breakdown products; artificial neural networks; omics data; metabolic response



Academic Editor: Ala F. Nassar

Received: 6 January 2026

Revised: 19 January 2026

Accepted: 20 January 2026

Published: 21 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The climate crisis is pushing industries to adopt sustainable practices focused on efficient resource use, waste reduction, and replacing fossil fuels with renewable energy. The chemical industries traditionally rely on oil and gas due to their availability and low cost [1]. However, as fossil fuel reserves deplete, and with the burgeoning energy demand, renewable sources like wind, solar, geothermal, hydropower, and biomass—an abundant

carbon source—are gaining much wider attention [2]. Among these renewable sources, biomass conversion into useful chemicals is a promising step towards sustainability. Of the 170 billion tons of renewable resources produced annually, only 3.5% is utilized by humans [3]. Lignin, a major component of lignocellulosic biomass, is increasingly seen as a valuable resource for producing fuels and chemicals [4]. Found in plant cell walls, lignin provides structural support and is abundant in wood and agricultural residues. Its complex molecular structure, comprising three phenylpropane units linked by ether and carbon–carbon bonds, along with hydroxyl and methoxy groups, makes it highly resistant to degradation [5]. This complexity presents a challenge in efficiently converting lignin into high-value products [6]. To overcome that, microbial degradation of lignin into desirable products can be an efficient alternate route.

Rhodopseudomonas palustris CGA009 is an alphaproteobacterium capable of thriving in diverse metabolic environments, including both phototrophic and chemotrophic conditions. It has the remarkable ability to fix carbon dioxide and nitrogen and can grow under either aerobic or anaerobic conditions. This versatility allows it to produce ATP using light, as well as organic or inorganic compounds for energy generation [7,8]. *R. palustris* was previously shown to catabolize *p*-coumarate, a lignin breakdown product (LBP) [9]. Thereby, it can be a promising bacterium to degrade different LBPs into value-added chemicals. However, to achieve that goal, a system-wide understanding of *R. palustris* is essential. A system-wide understanding of *R. palustris* is required to effectively apply it to converting lignin breakdown products (LBPs) into value-added chemicals because it allows us to identify and optimize the complex metabolic and regulatory process involved in LBP catabolism.

Genome-scale metabolic models (GSMs) are popular tools to gain systems-wide understanding of a living system [10,11]. So far, two GSMs of *R. palustris* have been reconstructed. The earlier GSM of *R. palustris* [12] was used to quantitatively understand trade-offs among a set of important biological objectives during different metabolic growth modes. Later, we developed another updated GSM of *R. palustris* (iRpa940), with an 84% accuracy, that successfully captured the relation between light-dependent energy production and the oxidation rate of quinol [13]. Moreover, based on iRpa940, we reconstructed the first-ever genome-scale metabolic and expression (ME-) model of *R. palustris* which successfully captured ferredoxin as a regulatory element in distributing electrons between carbon fixation and nitrogen fixation pathways, two major redox balancing pathways in *R. palustris* [14]. However, these first principles models usually connect metabolic reactions with bacterial growth rate. Thus, quantifying the relationship between non-metabolic genes/proteins and bacterial growth rate is not intuitive, as seen with enzymes like *boxB*, involved in anaerobic aromatic compound degradation, and *nuoF2*, which functions in the electron transport chain. In contrast, non-metabolic genes/proteins participate in processes like transport, signaling, stress response, and regulation, as exemplified by *rpa2624* (a sulfonate transport protein). However, conventional techniques such as differential expression analysis and pathway enrichment typically focus on individual genes or proteins with significant expression changes, potentially missing subtle, non-linear interactions and the combined effects of multiple factors. To resolve this issue, machine learning tools can be implemented. Previously, a machine learning (ML) tool was used to identify proteins that are associated with biomarkers in COVID-19 [15]. Moreover, ML was used to successfully identify proteins in hepatic carcinoma [16]. Furthermore, ML was used to predict antimicrobial proteins for numerous bacterial species [17]. Therefore, despite being a black box approach, ML can be a promising tool to identify proteins, metabolic and/or non-metabolic, associated with the bacterial growth rate.

We generated growth profiles and transcriptomic and proteomic datasets for *R. palustris* during catabolism of diverse lignin breakdown products (LBPs), including monolignols

(*p*-coumaryl alcohol, coniferyl alcohol, sinapyl alcohol), acid derivatives (*p*-coumarate and sodium ferulate), and kraft lignin, to investigate their metabolic routes. Using transcriptomic and proteomic features with growth rates as targets, we trained three machine learning models—Artificial Neural Networks (ANNs), Random Forest (RF), and Support Vector Machine (SVM). ANNs demonstrated the highest predictive performance, achieving 94% accuracy with transcriptomic data and 96% with proteomic data. To identify the genes and proteins directly associated with growth, the models were trained on the complete dataset following standard protocols [15]. Next, permutation feature importance analysis on the ANN models revealed the top twenty genes and proteins most impacting growth, and combining these insights identified eight key transport proteins (mostly associated with amino acid and sulfonate transportation) driving *R. palustris* growth on LBPs. Retraining the ANNs using these eight proteins yielded 86% accuracy for proteomics data and 76% for transcriptomics data. This work highlights the effectiveness of ANN models in predicting growth-associated genes and proteins, offering broader implications for optimizing microbial systems in bioconversion processes.

2. Materials and Methods

2.1. Growth Experiments of *R. palustris* for Different LBPs

The strain *R. palustris* BAA-98 CGA009 was sourced from the American Type Culture Collection (ATCC). All strains utilized in this study were preserved at -80°C . For storage, *R. palustris* strains were frozen in a 20% (*v/v*) glycerol solution, while *E. coli* strains were stored in 15% (*v/v*) glycerol. Upon recovery, *R. palustris* strains were cultured on solid 112 Van Niel's medium, and *E. coli* on LB agar (Miller, AMRESCO, Solon, OH, USA), both supplemented with the required antibiotics [18].

For the experimental setup, *R. palustris* seed cultures were initially grown aerobically and subsequently used for light-dependent biomass production (LBP) assays. Assays were performed in 50 mL of Photosynthetic Medium (PM) contained in 250 mL Erlenmeyer flasks, supplemented with 20 mM sodium acetate, 10 mM bicarbonate, and 15.2 mM ammonium sulfate. Aerobic seed cultures were diluted to an OD₆₆₀ of 0.2 (1/10th of the initial OD) and grown either aerobically in 50 mL PM within Erlenmeyer flasks or anaerobically in 13.5 mL PM within sealed 14 mL round-bottom tubes. Anaerobic cultures were illuminated during growth using an upper LED shelf light (SN-AG230-WIR-065) in an Algaetron incubator (Photon System Instruments, PSI). All cultures were maintained at 30°C with shaking at 275 rpm. For both aerobic and anaerobic conditions, 10 mM bicarbonate, 15.2 mM ammonium sulfate, and either 1 mM or 10 mM LBP were added as required. Each point in the growth curves represents the mean of three biological replicates. The cultivation duration for each lignin breakdown product was selected based on substrate-specific growth kinetics, as LBPs differ in bioavailability and metabolic complexity; cultures were therefore monitored for time intervals sufficient for capturing informative growth trends rather than until a uniform stationary phase was reached.

2.2. Transcriptomics Data Generation

Firstly, ribosomal RNA depletion was performed using NEBNext[®] rRNA Depletion Kit (Bacteria, New England Biolabs cat#E7860S). The rRNA-depleted RNA was purified by $2\times$ RNAClean XP beads (Beckman Coulter) and eluted in 45 μL of nuclease-free water. The purified RNA was then mixed with 4 μL of NEBNext First Strand Synthesis Reaction Buffer and 1 μL of random primers. The reaction was incubated at 94°C for 12 min for fragmentation. After incubation, the cultures were centrifuged to obtain cell pellets, which were then resuspended in RNeasy lysis buffer to prevent RNA degradation. Subsequently, to perform first-strand cDNA synthesis, 8 μL of NEBNext Strand Specificity Reagent and

2 μ L of NEBNext First Strand Synthesis Enzyme Mix was added to the reaction and the reaction was incubated at 25 °C 10 min, 42 °C 30 min, 70 °C 15 min. Second-strand reaction was then performed by adding 8 μ L of NEBNext Second Strand Synthesis Reaction Buffer with dUTP Mix (10 $\times$), 4 μ L of NEBNext Second Strand Synthesis Enzyme Mix, and 48 μ L of nuclease-free water. The reaction was incubated at 16 °C 1 h. The cDNA was purified by 1.8 $\times$ SPRIselect Beads (Beckman Coulter) and eluted in 50 μ L of nuclease-free water. The reaction was incubated at 16 °C 1 h. The cDNA was purified by 1.8 $\times$ SPRIselect Beads (Beckman Coulter) and eluted in 50 μ L of nuclease-free water. Subsequently, endprep reaction was performed by adding 7 μ L of NEBNext Ultra II End Prep Reaction Buffer and 3 μ L of NEBNext Ultra II End Prep Enzyme Mix into 50 μ L purified cDNA. Endprep reaction was incubated at 20 °C 30 min and 65 °C 20 min. Adaptor ligation reaction was then performed by adding 1 μ L of NEBNext Ligation Enhancer, 30 μ L of NEBNext Ultra II Ligation Master Mix, and 2.5 μ L of NEBNext Adaptor, diluted to 0.5 μ M in Adaptor Dilution Buffer. The mix was incubated at 20 °C 15 min. A total of 3 μ L of USER Enzymes (New England Biolabs) was then added to the ligation product and the reaction was incubated at 37 °C 15 min. The ligated product was purified by SPRIselect Beads (Beckman Coulter) and eluted in 15 μ L of nuclease-free water. PCR was carried out by adding 25 μ L of NEBNext Ultra II Q5 Master Mix, 5 μ L of i5 Primer, and 5 μ L of i7 Primer into 15 μ L of purified ligated product. PCR was performed at 98 °C 30 s, 15 cycles of 98 °C 10 s and 65 °C 75 s, and a final extension at 65 °C 5 min. The final library was then purified by SPRIselect Beads and loaded into Illumina NovaSeq6000 paired end 150 bp mode for sequencing.

Data quality control after obtaining the raw data (fastq files) and the quality of the original reads, including sequencing error rate distribution and GC content distribution, is evaluated using FastQC (v0.12.1, Cambridge, UK). The original sequencing sequences contain low quality reads and adapter sequences. To ensure the quality of data analysis, raw reads must be filtered to obtain clean reads, and the subsequent analysis is based on clean reads. Data filtering mainly includes the removal of adapter sequences in the reads, the removal of reads with high proportion of N (N denotes the unascertained base information), and the removal of low-quality reads. This process is carried out using fastp.

To remove ribosomal RNA (rRNA) from bacterial transcriptomic data, we utilized the SortMeRNA (v4.0, Lille, France) tool, a specialized local sequence alignment program known for its efficiency in filtering rRNA from high-throughput sequencing reads. Installed via Conda for assured compatibility, SortMeRNA employs an approximate seed-based algorithm that enables swift and sensitive rRNA identification and separation from the transcriptomic data.

The application of SortMeRNA resulted in the segregation of rRNA reads from non-rRNA reads, providing us with a refined dataset focused on coding regions and other functional genes. This targeted removal of rRNA components is pivotal for subsequent transcriptome assembly and functional gene analysis, ensuring that our research delves into the biologically relevant aspects of the transcriptome unencumbered by the predominance of rRNA sequences. The reference genome index was created by the build-index function in HISAT2 software package with default options. Then, the filtered clean reads were mapped to reference the genome by HISAT2, and the position and gene characteristics information were acquired. After the alignment, the generated SAM files were sorted to BAM files using samtools.

We used featureCounts (v2.0.2, Parkville, Australia) software from the subread package to quantify gene expression levels using mapped reads' positional information on the gene. The gene number in different expression levels, as well as the gene expression level of each single gene, was analyzed. DESeq was used to analyze the DEGs (differentially expressed genes) for samples with biological replicates and used edgeR for the samples

without replicates. During the analysis, samples should be grouped first so that the comparison between each two groups as a control–treatment pair can be performed later. During the process, Fold Change ≥ 1.5 and FDR < 0.05 are set as screening criteria.

2.3. Proteomics Data Generation

The cell pellets were lysed in Pierce RIPA buffer (Thermo Fischer Scientific, Waltham, MA, USA) containing 5 mM DTT and 1 $\times$ protease inhibitor (cOmplete EDTA-free protease inhibitor cocktail; Roche, Indianapolis, IN, USA) by shaking them at 95 °C for 10 min on a thermomixer (Eppendorf, Dallas, TX, USA). The samples were then centrifuged at 16,000 $\times g$ for 15 min, and the supernatants were transferred to a new tube. The proteins were assayed using the CB-X protein assay (G-Biosciences, St. Louis, MO, USA). A total of 50 μ g of reduced protein was alkylated with 20 mM iodoacetamide for 40 min and quenched with DTT. The proteins were then precipitated with acetone, and the pellets washed 3 times with 70% ethanol. Proteins were resuspended in 50 μ L of 50 mM Tris/HCl, pH 8.0 containing 1 μ g Lys-C and digested for 4 h, followed by further digestion with 1 μ g trypsin overnight at 37 °C. A quality control reference sample was prepared by mixing all the samples with the same 1:1 ratio to run between every 16 samples to check for instrument performance deviation. The sequence order of the samples was randomized using block randomization.

Each digest was run by nano liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) using an Ultimate 3000 RSLCnano system coupled to an Orbitrap Eclipse mass spectrometer (Thermo Fisher Scientific). Briefly, peptides were first trapped and washed on a trap column (Acclaim PepMapTM 100, 75 μ m $\times$ 2 cm, Thermo Fisher Scientific). Separation was then performed on a C18 nano column (Acquity UPLC[®] M-class, Peptide CSHTM 130A, 1.7 μ m 75 μ m $\times$ 250 mm, Waters Corp, Milford, MA, USA) at 300 nL/min with a gradient from 5 to 22% over 75 min. The LC aqueous mobile phase was 0.1% (*v/v*) formic acid in water, and the organic mobile phase was 0.1% (*v/v*) formic acid in 100% (*v/v*) acetonitrile. Mass spectra were acquired using the data-dependent mode with a mass range of *m/z* 375–1500, resolution 120,000, AGC (automatic gain control) target 4×10^5 , and maximum injection time 50 ms for the MS1. Data-dependent MS2 spectra were acquired by HCD in the ion trap with a normalized collision energy (NCE) set at 30%, AGC target set to 5×10^4 , and a maximum injection time of 86 ms.

The identification and quantitation of the proteins were performed using the Proteome Discoverer (Version 2.4; Thermo Fisher Scientific) utilizing the MASCOT search engine (Version 2.7.0; Matrix Science Ltd., London, UK). The search was performed against an in-house modified version of the cRAP database (theGPM.org/cRAP) and the *Rhodopseudomonas palustris* (version_20230110) database obtained from UniProt (ID: UP000001426_258594, www.uniprot.org), assuming the digestion enzyme trypsin and a maximum of 2 missed cleavages. Mascot was searched with a fragment ion mass tolerance of 0.06 Da and a parent ion tolerance of 15.0 ppm. Deamidation of asparagine and glutamine and oxidation of methionine were specified in Mascot as variable modifications, while the carbamidomethylation of cysteine was fixed. Peptides were validated by a Percolator with a 0.01 posterior error probability (PEP) threshold. The data were searched using a decoy database to set the false discovery rate (FDR) to 1% (high confidence). Only proteins identified with a minimum of 2 unique peptides and 5 peptide-spectrum matches (PSMs) were further analyzed for quantitative changes. The peptides were quantified using the precursor abundance based on intensity. The peak abundance was normalized using the total peptide amount. The peptide group abundances are summed for each sample and the maximum sum for all files is determined. The normalization factor used is the factor of the sum of the sample and

the maximum sum in all files. The normalized abundances were scaled for each protein so that the average abundance is 100.

2.4. Different ML Algorithms

Three distinct machine learning models—Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Regression (SVR)—were trained using SciPy (v1.17.0) to predict microbial growth rates using a high-dimensional proteomics dataset comprising 1854 features. The dataset was loaded into a pandas DataFrame.

Frame and feature standardization was performed using StandardScaler (v1.8.0) to ensure zero mean and unit variance, a critical preprocessing step when dealing with models sensitive to feature scaling. The ANN model was implemented as a deep Multi-Layer Perceptron (MLP) with 12 hidden layers, each containing 200 neurons, resulting in a highly parameterized model. The training process was carried out over 1000 iterations (epochs) using a random seed of 42 for reproducibility. To further interpret the model, permutation feature importance was computed over 30 iterations, which allowed us to assess the robustness and consistency of feature rankings. By repeatedly shuffling each feature and measuring the resulting decrease in predictive accuracy, this approach mitigates the impact of random variation and highlights features that consistently contribute to model performance across iterations. From this analysis, we identified the top twenty proteins most predictive of the growth rate (ensuring a manageable set of features for comprehensive functional annotation and pathway analysis without overwhelming the biological interpretation process), providing a stable and reliable set of critical biological markers directly associated with *R. palustris* growth during LBP catabolism. The Random Forest (RF) regressor was implemented with 1000 estimators (decision trees), a model configuration designed to capture complex interactions between features while reducing variance through ensemble learning. The model's inherent feature importance mechanism was used to assess the contribution of individual proteomic features, further aiding in the biological interpretation of the data. For the Support Vector Regression (SV) model, a radial basis function (RBF) kernel was selected to capture nonlinear relationships within the proteomic data. Hyperparameters were tuned with a regularization parameter (C) set to 100 and the epsilon set to 0.01, optimizing the model's balance between bias and variance.

2.5. Statistical Analysis for Top Genes and Proteins

Statistical analyses on the nine statistically significantly abundant proteins (between aerobic and anaerobic conditions) were performed to identify the proteins with highest expression variability. For each protein of those nine proteins, the mean expression level and standard deviation were calculated across the fourteen conditions. To normalize the variability across proteins, the coefficient of variation (CV) was calculated as the ratio of the standard deviation to the mean expression level. The proteins were then ranked according to their CV values, with lower CV values indicating more consistent expression levels between aerobic and anaerobic conditions. Conversely, proteins with higher CV values exhibited greater variability between aerobic and anaerobic conditions.

2.6. Simulation Platform

ANN, RF, SV, and CV algorithms were implemented in Python using an Intel(R) Core(TM) i5-8250U CPU 1.60 GHz HP laptop with 8.00 GB of RAM and 64-bit operation with a Windows 11 Home operating system.

3. Results and Discussion

3.1. Growth of *R. palustris* on Different Lignin Breakdown Products

Although the ability of *R. palustris* to catabolize *p*-coumarate is already established [9], in this work, we wanted to reassess its capability to catabolize other LBPs as well. Wild-type cultures were supplemented with various LBPs, including kraft lignin. Specifically, the monolignols *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol, along with two acid derivatives, *p*-coumarate and sodium ferulate, were evaluated as potential carbon sources. These substrates were selected based on their prevalence in the depolymerization of kraft lignin. Acetate, a carbon source with a simple catabolic route [13], was included as a positive control. Growth curves were generated for each substrate under both aerobic and anaerobic conditions (Supplementary Figure S1), and growth rates were calculated using logistics model.

Notably, *R. palustris* is unable to grow on certain LBPs when these are the sole carbon sources under varying oxic conditions, corroborating findings from previous studies [19,20]. In such cases, 10 mM of sodium acetate was supplemented as an additional carbon source providing energy (ATP) through the TCA cycle to support the energy-intensive enzymatic processes required for LBP catabolism. However, to confirm that *R. palustris* is utilizing the LBPs for biomass production rather than relying on acetate, the OD660 of these cultures must have exceeded that of the positive acetate control. Among the substrates tested, only *p*-coumarate was consistently consumed by *R. palustris* without the need for acetate supplementation, under both aerobic and anaerobic conditions. In contrast, *p*-coumaryl alcohol and the methoxylated monomers were more resistant to degradation. We hypothesize that acetate, which directly enters the TCA cycle, efficiently generates ATP to support the energy-intensive catabolism of lignin breakdown products (LBPs), providing the necessary energy to drive their enzymatic degradation.

Building on the initial experiments, which provided clear evidence of *R. palustris*' ability to catabolize these LBPs, we further explored the organism's transcriptomic and proteomic response when metabolizing these substrates. For transcriptomic analysis, two biological replicates were harvested during the mid-exponential growth phase (Supplementary Table S1) and submitted for Next-Generation Sequencing (NGS). Additionally, five biological replicates for each condition were collected to analyze the proteome profiles. We also wanted to make sure that photosynthetic machinery was active in the anoxic growth. One key regulator, *fixK*, was proportionally upregulated in our anaerobic samples compared to aerobic samples, suggesting its role in adapting to anoxic conditions. Additionally, we observed a reduced abundance of genes encoding oxygen-dependent enzymes such as *hemF* (RPA1514), cytochrome bd oxidases (RPA1319, RPA4452, RPA4793-RPA4794), and cytochrome aa3 oxidases (RPA1453, RPA4183, RPA0831-RPA0836) (Table S4). In contrast, genes associated with high-affinity cytochrome cbb3 oxidase (RPA0015-RPA0019), oxygen-independent coproporphyrinogen oxidase HemN (RPA1666), and components of the photosynthetic apparatus (RPA1505-RPA1507, RPA1521-RPA1548, RPA1667-RPA1668, RPA3568) were upregulated. These findings indicate a strong activation of photosynthetic machinery and related pathways under anoxic, light-exposed conditions, consistent with *R. palustris* physiology. This system-wide shift in metabolism, redox balance, and pigment biosynthesis confirms that our photoheterotrophic cultures were indeed anaerobic, as also validated by *FixK*-regulated gene induction (RPA1006-RPA1007, RPA1554). These observations inform our analysis by demonstrating how anaerobic growth in the presence of light drives specific metabolic and regulatory adaptations. By training different machine learning (ML) models where transcriptomics/proteomics data are featured and the growth rates are targeted, we can identify important genes/proteins associated with the *R. palustris*

growth under aerobic and anaerobic conditions for different LBPs. The overall workflow of the work can be found in Figure 1a,b.

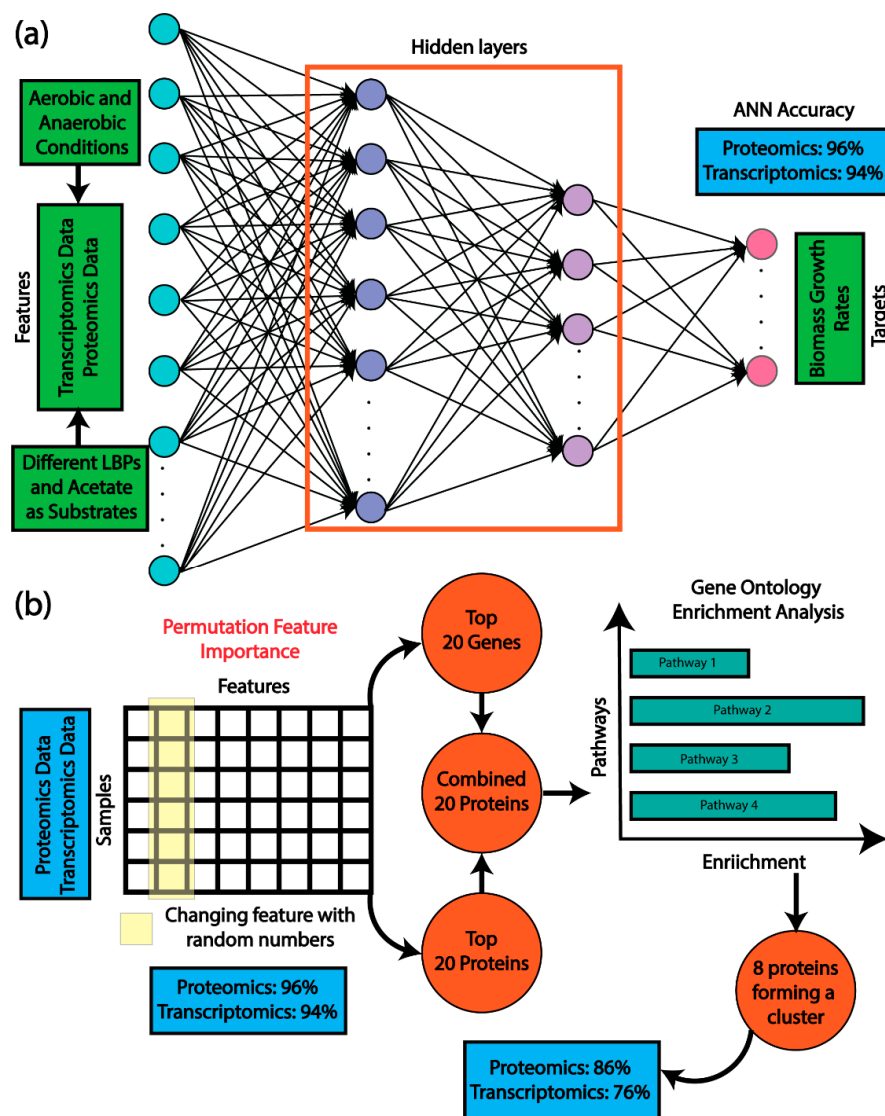


Figure 1. Artificial Neural Network (ANN) predicts the top 20 growth-associated genes and proteins. (a) ANN predicted the growth rates from transcriptomics and proteomics data with very high accuracy. (b) Later, permutation feature importance was used to determine the top twenty genes and proteins, affecting the growth rates most. Furthermore, a list of top twenty proteins, determined from the combined list of top twenty genes and top twenty proteins, yielded eight proteins through gene set enrichment analysis, and further training the ANN with those eight proteins and genes only resulted in 86% and 76% accuracy, respectively.

3.2. Artificial Neural Networks Accurately Predicted Growth Rates from ‘Omics’ Data

Once the transcriptomics and proteomics data were generated, these datasets were used to build different ML algorithms. As proteins are often difficult to detect compared to the mRNA, for training ML algorithms, we only kept the genes in the transcriptomics dataset that matched with the detected proteins. Furthermore, we normalized both the proteomics and the transcriptomics datasets by subtracting the mean and scaling it to the unit variance. In the process, we were able to keep 1855 genes and 1855 proteins to train ML algorithms for fourteen different conditions.

Next, for both the transcriptomics and the proteomics datasets, we benchmarked three different machine learning models—Artificial Neural Network (ANN), Random

Forest algorithm (RF), and Support Vector Machine (SV) to assess the suitability of each for predicting growth rates based on transcriptomics and proteomics data individually. In these algorithms, proteomics/transcriptomics data were the features and growth rates were the target. As the primary objective of this study is to pinpoint the proteins and genes that exert the greatest influence on the growth of *R. palustris* during LBP catabolism, we adopted an approach in which the full transcriptomic and proteomic datasets were used to train each machine learning model. This strategy ensured that no potentially growth-relevant molecular features were excluded from analysis and was designed to generate a comprehensive and interpretable catalog of growth-associated features by quantifying the contribution of each gene and protein to growth prediction, consistent with methodologies applied in prior work [15]. In that way, the entire genome was used to train the ML models against growth rates, which is more realistic than using only a part of the genome to train ML models. Among these three algorithms, ANN achieved the least mean absolute error (MAE) for both transcriptomics and proteomics data (0.00129 and 0.00111, respectively), followed by RF (0.00547 and 0.00524, respectively) and SV (0.00859 and 0.00841) (Figure 2a). Similarly, ANN achieved the least mean square error (MSE) for both transcriptomics and proteomics data (2.76×10^{-6} and 2.19×10^{-6} , respectively), followed by RF (3.69×10^{-5} and 3.45×10^{-5} , respectively) and SV (8.23×10^{-5} and 7.92×10^{-5}) (Figure 2b). In terms of accuracy, ANN showed much higher accuracy in predicting the growth rates from transcriptomics and proteomics data (93.68% and 95.99%, respectively) compared to the RF (76.28% and 77.29%, respectively) and SV (65.89% and 66.32%, respectively) (Figure 2c).

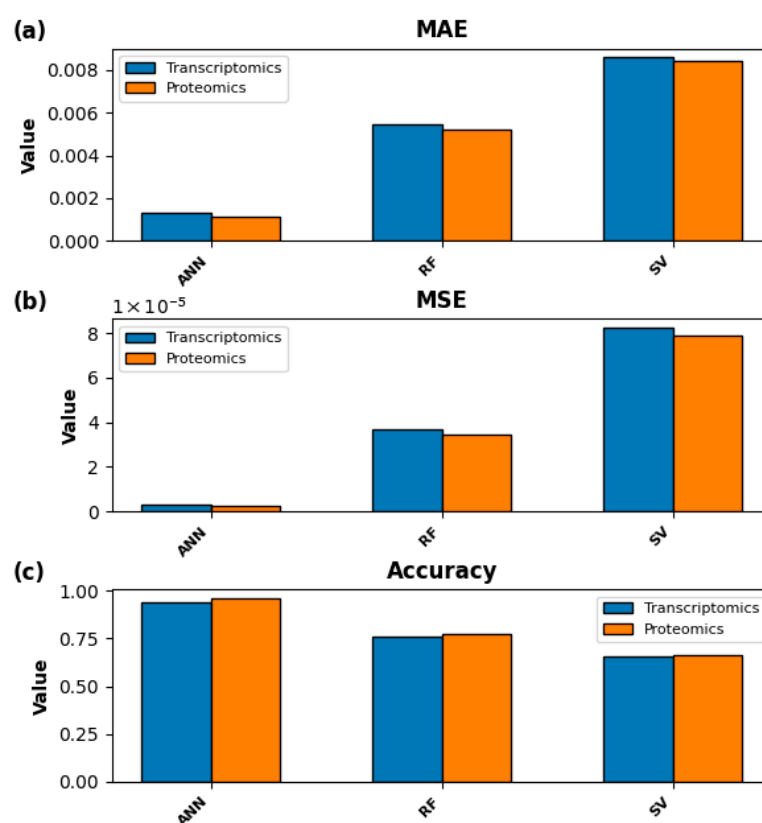


Figure 2. Artificial Neural Network (ANN) is the best-performing machine learning algorithm compared to the Random Forest (RF) algorithm and Support Vector (SV) Machine in *R. palustris*. (a) Mean Absolute Error (MAE) for three different machine learning algorithms. (b) Mean Squared Error (MSE) for all three different machine learning algorithms. (c) Accuracy for three different machine learning algorithms.

Overall, from the training of three different machine learning algorithms, ANN was the best-performing algorithm for capturing growth rates from transcriptomics and proteomics data. Moreover, proteomics data are a better indicator of enzymatic activity compared to the mRNA, as the latter one undergoes fast dilution and degradation due to its unstable nature [21].

3.3. Permutation Feature Importance Identified Top Twenty Growth-Associated Genes

As ANN was the best performing machine learning algorithm, we chose ANN to find the most important genes and proteins that impacted the growth rates. At first, from the ANN algorithm, we generated the top twenty genes impacting the growth rates using permutation feature importance from all the given conditions [22] (Figure 3a). Permutation feature importance is an algorithm used to evaluate the importance of features in an ML model by measuring how performance decreases when a feature's values are randomly shuffled. The underlying idea is that if a feature is important for making accurate predictions, randomly shuffling its values should lead to a significant drop in model performance. If shuffling a feature does not affect performance, the model does not rely on that feature to make predictions. The expression pattern of those twenty genes can be observed in Figure 3b.

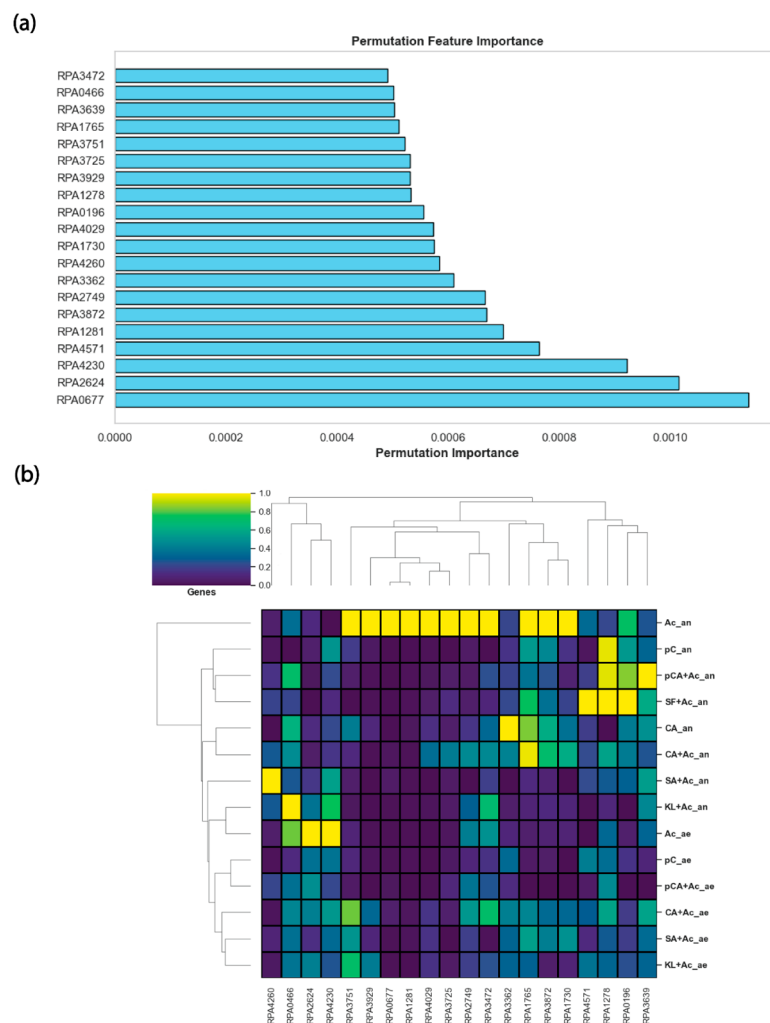


Figure 3. ANN predicted the top twenty genes that affected the growth rate most. (a) The top twenty growth-associated genes from the deep learning framework using permutation feature importance. (b) Top twenty growth-associated gene expression profiles.

From the top twenty genes (Supplementary Table S2), there were only four genes that were directly associated with metabolism: *rpa0677* (*boxB*), *rpa4260* (*nuoF2*), *rpa1765* (*ech*), and *rpa3472* (*ilvD1*). We determined the metabolic activity of a given gene based on KEGG database. For example, if a gene is associated with a metabolic reaction, that gene is considered to be directly associated with metabolism. Benzoyl-CoA 2,3-epoxidase (*boxB*) plays a key role in the anaerobic degradation of aromatic compounds, particularly benzoate, in *R. palustris*. This enzyme is part of the pathway that enables *R. palustris* to utilize aromatic compounds as a carbon and energy source under anaerobic conditions. *boxB* catalyzes the epoxidation of benzoyl-CoA, forming 2,3-epoxybenzoyl-CoA, which is a crucial step in the ring cleavage and degradation of the aromatic structure, thereby may play a role in the degradation of many other LBPs [23]. Moreover, *boxB* often works with a redox partner (e.g., flavoprotein reductase) to receive electrons and enable the epoxidation reaction, utilizing reduced electron carriers such as NADPH or ferredoxin [24]. Next, NADH-quinone oxidoreductase (*nuoF2*) in *R. palustris* plays a key role in its electron transport chain by transferring electrons from NADH to quinones, such as ubiquinone, while simultaneously pumping protons across the membrane. This creates a proton gradient used to generate ATP, crucial for the cell's energy production. In *R. palustris*, which thrives in diverse metabolic conditions [8], *nuoF2* helps maintain redox balance, adapting to various electron acceptors and energy sources [25]. Enoyl-CoA hydratase (*ech*) plays a key role in the β -oxidation of fatty acids, catalyzing the hydration of enoyl-CoA to 3-hydroxyacyl-CoA. This is essential for fatty acid degradation, providing energy and carbon for the cell; additionally, it may be involved in the degradation of hydrocarbons, helping to metabolize alkenes and alkanes [26]. Finally, xylonate dehydratase (*xyID*) plays a crucial role in the metabolism of pentose sugars, specifically in the conversion of xylonate to 2-keto-3-deoxyxylonate, a key intermediate in the pentose and glucuronate interconversion pathways [27]. While *R. palustris* cannot directly catabolize xylose or glucose, this enzyme may facilitate the utilization of related metabolites derived from plant-derived polysaccharides, supporting its adaptation to lignocellulose-rich environments.

Next, from gene ontology analysis of those twenty genes using ShinyGO (v0.85) [28], we found only one cluster of eight genes, associated with the signaling and transportation mechanism of different substrates. These genes are *rpa0466* (metalloprotease inhibitor), *rpa2624* (sulfonate transport), *rpa3362* (hypothetical protein), *rpa3725* (ABC transporter substrate-binding protein), *rpa3751* (cache domain-containing protein), *rpa3872* (hypothetical protein), *rpa4029* (ABC transporter substrate-binding protein), and *rpa4571* (BA14K family protein). In *R. palustris*, the listed genes represent a diverse array of functions that contribute to the organism's metabolic versatility and environmental adaptability. *rpa0466*, a metalloprotease inhibitor, likely plays a regulatory role in proteolysis, protecting proteins from degradation under certain conditions [29]. *rpa2624* is involved in sulfonate transport, enabling the organism to utilize sulfur from organic sources, which is important in sulfur-limited environments [30]. Both *rpa3725* and *rpa4029* are substrate-binding proteins of ABC transporters, which facilitate the uptake of various substrates critical for nutrient acquisition [31]. *rpa3751*, a cache domain-containing protein, is likely involved in sensing extracellular signals or nutrients, guiding cellular responses [32]. *rpa3362* and *rpa3872*, both hypothetical proteins, might represent uncharacterized functions in metabolism or stress response. Lastly, *rpa4571*, a BA14K family protein, could have a role in stress resistance or environmental interactions [33]. Together, these genes reflect the complex regulatory, transport, and environmental sensing capabilities of *R. palustris*, enhancing its ability to thrive in diverse ecosystems. Comparing the result with gene ontology analysis, StringDB database could not predict any interaction among the proteins encoded by these genes. The lack of predicted interactions in StringDB, along with the high prevalence of hypothetical

and unannotated genes in *R. palustris*, underscores the unique and relatively understudied nature of this organism.

The rest of the genes, such as *rpa4230*, *rpa1281*, *rpa2749*, *rpa1730*, *rpa0196*, *rpa1278*, *rpa3929*, and *rpa3639*, did not fall into specific categories in gene ontology analysis, with *rpa4230* being identified as a gene with hypothetical function. Details of the rest of the genes can be accessed in the Supplementary Table S2.

3.4. Permutation Feature Importance Identified Top Twenty Growth-Associated Proteins

Similarly to the top twenty genes, we also identified the top twenty growth-associated proteins from all the given conditions (Supplementary Table S3) using permutation feature importance from the ANN model trained using proteomics data (Figure 4a). Their abundance profile can be observed in Figure 4b. Interestingly, there is no overlap between the list of top twenty genes and the top twenty proteins, thus indicating weak correlation between transcriptomics and proteomics data. The lack of overlap between transcriptome and proteome datasets indicates a weak correlation between mRNA and protein levels, a common phenomenon in alphaproteobacteria like *Rhodobacter sphaeroides* [34]. This suggests significant post-transcriptional regulation, where factors such as mRNA stability, translation efficiency, and protein turnover decouple transcript and protein abundances. Among those proteins, three were associated with metabolism: RPA0069 (tryptophan synthase subunit beta), RPA0532 (beta-ketoacyl-ACP reductase), and RPA2720 (glutathione-dependent disulfide-bond oxidoreductase). In *R. palustris*, the tryptophan synthase subunit beta plays a vital role in the biosynthesis of the essential amino acid tryptophan. This enzyme catalyzes the final step in the tryptophan biosynthesis pathway, converting indole and serine into tryptophan [35]. As tryptophan is a precursor for several important biomolecules, including proteins and signaling compounds, its synthesis is critical for *R. palustris* to maintain cellular function and adapt to various environmental conditions. The ability to synthesize tryptophan internally allows *R. palustris* to thrive in nutrient-limited environments where external sources of amino acids may be scarce, contributing to its metabolic independence and ecological versatility. Beta-ketoacyl-ACP reductase is a key enzyme in the fatty acid biosynthesis pathway. It catalyzes the reduction of beta-ketoacyl-ACP to beta-hydroxyacyl-ACP, an essential step in the elongation cycle of fatty acid synthesis. These fatty acids are crucial components of the cell membrane and are also used for energy storage. By facilitating the production of lipids, beta-ketoacyl-ACP reductase plays a critical role in maintaining membrane integrity, enabling adaptation to environmental stress, and supporting energy metabolism [36]. This enzyme's function is fundamental to the organism's ability to build cellular structures and thrive in diverse ecological niches. Finally, glutathione-dependent disulfide-bond oxidoreductase plays a crucial role in maintaining cellular redox balance by catalyzing the reduction in disulfide bonds in proteins, using glutathione as a cofactor [37]. This enzyme is essential for protecting the cell against oxidative stress by facilitating proper protein folding, repair, and regeneration under adverse conditions. By reducing oxidized protein thiols, it helps mitigate the damage caused by reactive oxygen species, thereby preserving protein function and structural integrity. Additionally, it supports broader cellular defense mechanisms against environmental fluctuations, enhancing the organism's adaptability. By maintaining redox homeostasis, glutathione-dependent disulfide-bond oxidoreductase enables *R. palustris* to efficiently manage oxidative stress, sustain metabolic functions, and thrive in dynamic environments where redox imbalances could compromise cell viability.

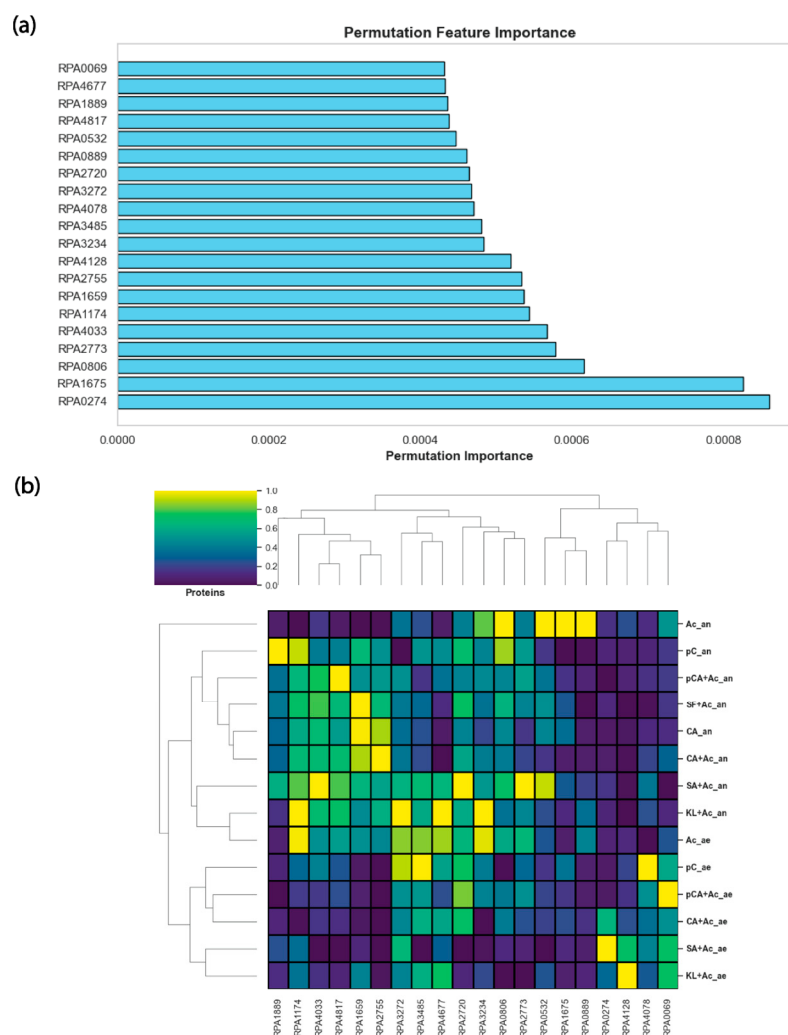


Figure 4. ANN predicted the top twenty proteins that affected the growth rate the most. (a) The top twenty growth-associated proteins from the deep learning framework using permutation feature importance. (b) Top twenty growth-associated proteins' abundance profile.

Next, the StringDB network reveals that only RPA3272 (large subunit ribosomal protein L1) and RPA3234 (50S ribosomal protein L18) out of the other proteins were strongly connected with each other among those twenty proteins (Supplementary Figure S2). In addition, a previous gene essentiality study found both proteins as essential [38]. In *R. palustris*, RPA3272 and RPA3234 are essential components of the ribosome, which plays a critical role in protein synthesis. RPA3272 is involved in binding rRNA and is important for the release of tRNA during translation, ensuring efficient protein production [39]. RPA3234 helps stabilize the structure of the 50S ribosomal subunit by binding to 5S rRNA, which is crucial for the assembly and function of the ribosome [39]. These ribosomal proteins ensure the accurate and efficient synthesis of enzymes and proteins that drive key metabolic pathways, including photosynthesis, nitrogen fixation, and carbon utilization. By supporting robust protein translation, RPA3272 and RPA3234 contribute to the metabolic flexibility and environmental adaptability of *R. palustris*.

While the rest of the proteins, such as RPA4677, RPA1889, RPA4817, RPA0889, RPA4078, RPA3485, RPA4128, RPA2755, RPA1659, RPA1174, RPA4033, RPA0806, RPA1675, and RPA0274, did not fall into specific categories in gene ontology/StringDB analysis, we explored their role on an individual basis. The caspase family protein (RPA4677) could be involved in programmed cell death or stress responses, playing a role in cellular regulation [40]. Domain-containing proteins such as DUF2019 (RPA1889) and DUF2188

(RPA4128) suggest the presence of proteins with unidentified functions potentially linked to regulatory or structural roles within the cell. The response regulator transcription factor (RPA4817) may participate in signal transduction pathways, enabling *R. palustris* to respond to environmental stimuli [41]. Proteins like the molecular chaperone (RPA0889) and DNA starvation/stationary phase protection protein (RPA2755) are likely involved in stress protection, helping the bacterium survive adverse conditions. Enzymes such as CoA transferase (RPA3485), uroporphyrinogen-III synthase (RPA4033), and 23S rRNA methyltransferase (RPA1174) point to key roles in metabolic pathways and gene expression. Proteins related to transport and chemotaxis, such as the ABC transporter substrate-binding protein (RPA0806), the methyl-accepting chemotaxis protein (RPA1675), and the P-II family nitrogen regulator (RPA0274), are likely crucial for nutrient uptake and environmental navigation, reflecting the bacterium's ability to thrive in diverse habitats. Overall, these proteins highlight the complex regulatory and metabolic network that underpins the survival and versatility of *R. palustris*. Details of the rest of the proteins can be accessed in the Supplementary Table S3.

3.5. Statistical Analysis Revealed Regulatory and Conserved Proteins During Lignin Breakdown

Since these top twenty proteins have diverse functions, we also wanted to investigate how their functionality varies with the availability of oxygen. Thus, we determined how many of the top twenty most important proteins showed statistically significant differences in abundance between aerobic and anaerobic conditions. Of these top twenty proteins, we found that nine exhibited statistically significant differences: RPA0069, RPA1889, RPA4817, RPA4078, RPA4128, RPA2755, RPA1659, RPA1174, and RPA4033 (Figure 5).

Differentially Abundant Proteins

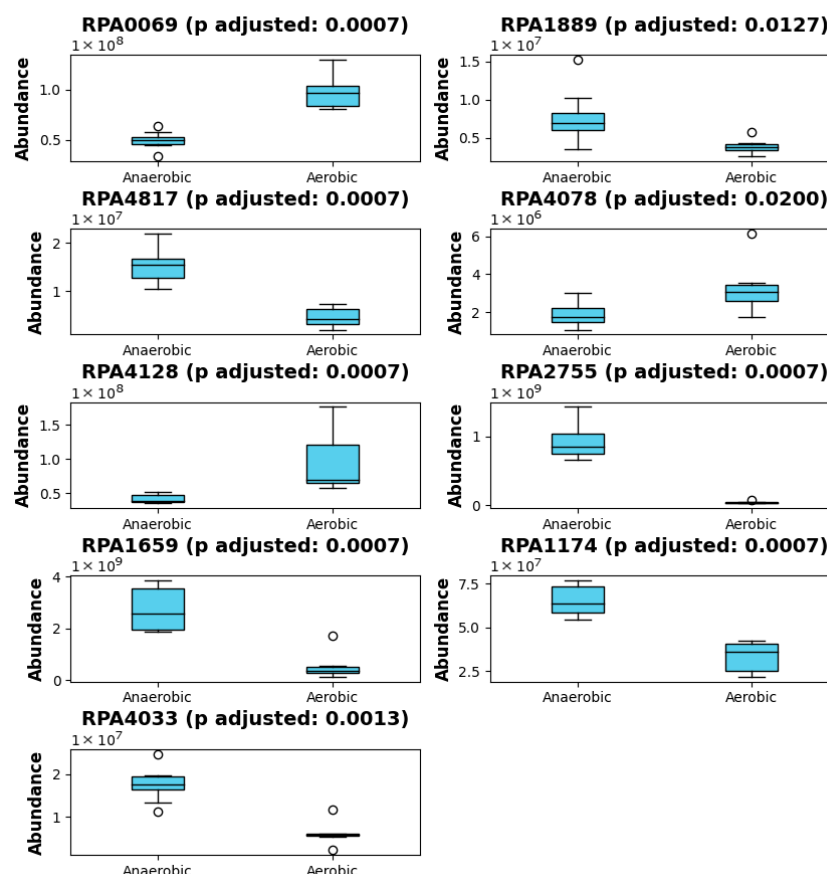


Figure 5. Out of top twenty growth-associated proteins, nine are associated with aerobic conditions.

The differential abundance of these proteins between aerobic and anaerobic conditions suggests that these proteins play distinct roles in cellular responses to oxygen availability. For instance, tryptophan synthase subunit beta (RPA0069) may be involved in metabolic pathways that are more active under oxygen-rich environments, where biosynthetic processes like amino acid synthesis require higher energy. The response regulator transcription factor (RPA4817) could indicate alterations in gene regulation in response to changes in redox conditions. Proteins such as the DNA starvation/stationary phase protection protein (RPA2755) and hemerythrin domain-containing protein (RPA1659) are likely involved in the cell's adaptive response to stress, where the lack of oxygen imposes survival challenges, prompting the activation of protective or metabolic adjustments. The presence of DUF domain-containing proteins (e.g., RPA1889 and RPA4128), which have unknown or poorly characterized functions, might represent novel pathways or processes important in anaerobic adaptation. Moreover, the activity of enzymes such as 23S rRNA methyltransferase (RPA1174) and uroporphyrinogen-III synthase (RPA4033) hints at modifications in translation machinery and heme biosynthesis, respectively, both of which are crucial under varying oxygen levels, potentially impacting cellular respiration and energy production. Finally, the appearance of a hypothetical protein (RPA4078) highlights that there might still be unknown factors contributing to the organism's response to oxygen tension.

Among those nine proteins, whose abundances changed significantly between aerobic and anaerobic conditions, we also wanted to evaluate which proteins had the highest differences in abundances. To accomplish that, we calculated the coefficient of variation (CV) for all nine proteins across aerobic and anaerobic conditions (Supplementary Figure S3). The higher the CV is, the more differentially abundant the protein is between aerobic and anaerobic conditions. From the CV analysis, RPA2755 (DNA starvation/stationary phase protection protein) showed the highest CV between aerobic and anaerobic conditions. RPA1174 (23S rRNA methyltransferase) showed the least CV among those nine proteins.

The remaining eleven (RPA4677, RPA0532, RPA0889, RPA2720, RPA3272, RPA3485, RPA3234, RPA2773, RPA0806, RPA1675, and RPA0274) out of those top twenty proteins did not exhibit statistically significant differences in their abundances between aerobic and anaerobic conditions. This lack of differential abundance suggests that their roles are not strongly influenced by the availability of oxygen. For example, RPA4677, a caspase family protein, may participate in processes like programmed cell death, but oxygen levels do not appear to impact its regulation. Similarly, RPA0532, a beta-ketoacyl-ACP reductase involved in fatty acid synthesis, and RPA0889, a molecular chaperone which assists in protein folding, remain unaffected by shifts in oxygen levels. Even proteins with roles in redox balance, such as RPA2720, a glutathione-dependent disulfide-bond oxidoreductase, and RPA3485, a CoA transferase linked to metabolic processes, show no significant response to oxygen conditions. Ribosomal proteins like RPA3272 and RPA3234, essential for protein synthesis, as well as RPA0806, an ABC transporter substrate-binding protein, and RPA1675, a methyl-accepting chemotaxis protein, are similarly unaffected. The hypothetical protein RPA2773 and the P-II family nitrogen regulator RPA0274 also remain stable regardless of oxygen availability. This suggests that while these proteins play essential roles in the cell, their expression is not directly modulated by aerobic or anaerobic growth, indicating their functions may be more constant across different oxygen environments.

3.6. Combined Permutation Feature Importance Highlights Amino Acid Transporters' Role in LBP

As encoding proteins is a process starting from mRNA, we thereby combined the permutation feature importance score for both the transcriptomics and the proteomics datasets and came up with a combined permutation feature importance for each of the proteins. The list can be accessed in the Supplementary Table S4. From the list, similarly to

the previous section, we identified the top twenty proteins. From the list of these twenty proteins, five came from the list of top twenty proteins (RPA4677, RPA0806, RPA1675, RPA0274, and RPA0532) and ten came from the list of top twenty genes (RPA2624, RPA4571, RPA3362, RPA3872, RPA0677, RPA4230, RPA0196, RPA1281, RPA2749, and RPA3929). Interestingly, five came from outside the list of top twenty proteins and genes (RPA4797, RPA1428, RPA2455, RPA1190, and RPA4297), showcasing the importance of combining transcriptomics and proteomics data to predict the growth rates of *R. palustris* under different LBPs.

Among those five proteins, which came outside of the list of top twenty proteins and genes, 7-carboxy-7-deazaguanine synthase (RPA1190) and aldo/keto reductase (RPA4297) were the metabolic proteins. In *R. palustris*, both RPA1190 and RPA4297 play important roles in catabolizing LBPs under both aerobic and anaerobic conditions. RPA1190 is involved in the biosynthesis of 7-deazapurines, which can influence metabolic pathways associated with lignin-derived aromatic compound processing [42]. Its activity might be crucial for cofactor biosynthesis that aids in the degradation of complex lignin structures. On the other hand, RPA4297, an aldo/keto reductase, catalyzes the reduction of aldehydes and ketones, which are common intermediates in lignin depolymerization. Under aerobic conditions, it helps detoxify reactive oxygen species generated during lignin degradation, while in anaerobic conditions, it assists in reducing aromatic aldehyde intermediates to less toxic alcohols, facilitating the complete mineralization of lignin-derived compounds [43]. Three other proteins were also predicted outside of the list of top twenty genes and proteins. Proteins such as the amino acid ABC transporter substrate-binding protein (RPA4797), the MetQ/NlpA family ABC transporter substrate-binding protein (RPA1428), and the M23 family metallopeptidase (RPA2455) are vital for nutrient acquisition and environmental adaptation during lignin breakdown. RPA4797 plays a key role in importing amino acids, which are crucial for the cellular growth and enzyme production necessary for lignin degradation pathways [44]. Similarly, RPA1428, a member of the MetQ/NlpA family protein, is involved in binding substrates such as sulfur-containing compounds like methionine, which can be essential for cellular redox [45] balance during lignin depolymerization. RPA2455, a metallopeptidase from the M23 family, facilitates the breakdown of peptide bonds in proteins or small peptides that may accumulate from microbial community interactions or host biomass degradation [29]. By processing and recycling these molecules, these proteins help *R. palustris* efficiently metabolize lignin-derived compounds and thrive under varying environmental conditions, optimizing nutrient acquisition and waste processing. Together, these proteins contribute to the efficient utilization of lignin as a carbon source, allowing *R. palustris* to thrive in diverse environments.

Next, we again performed gene ontology analysis using ShinyGO and protein–protein interaction using StringDB. StringDB returns no interactions among all these proteins. However, gene ontology analysis identified eight amino acid transport proteins (RPA4797, RPA1428, RPA4677, RPA0806, RPA2624, RPA4571, RPA3362, and RPA3872), associated with signaling, forming a cluster (Figure 6a). Furthermore, only these eight proteins were used to train the ANN algorithm again, which resulted in 86% accuracy (Figure 6b). We also used the corresponding eight genes to train the ANN algorithm again, which resulted in 78% accuracy (Figure 6b). Similarly to the previous cases, proteomics-data-based predictions of growth rate performed better compared to the transcriptomics-data-based growth predictions using ANN. Permutation feature importance again ranked these transport proteins from the most important to the least. RPA3362, a hypothetical protein, was found to be the most important protein dictating the growth, and RPA4677, a caspase protein, was identified to be the least important among those eight proteins (Figure 6c). Thus, by combining permutation feature importance for both the transcriptomics and

proteomics datasets, we identified eight proteins which were able to predict the growth rates of *R. palustris* with an excellent accuracy.

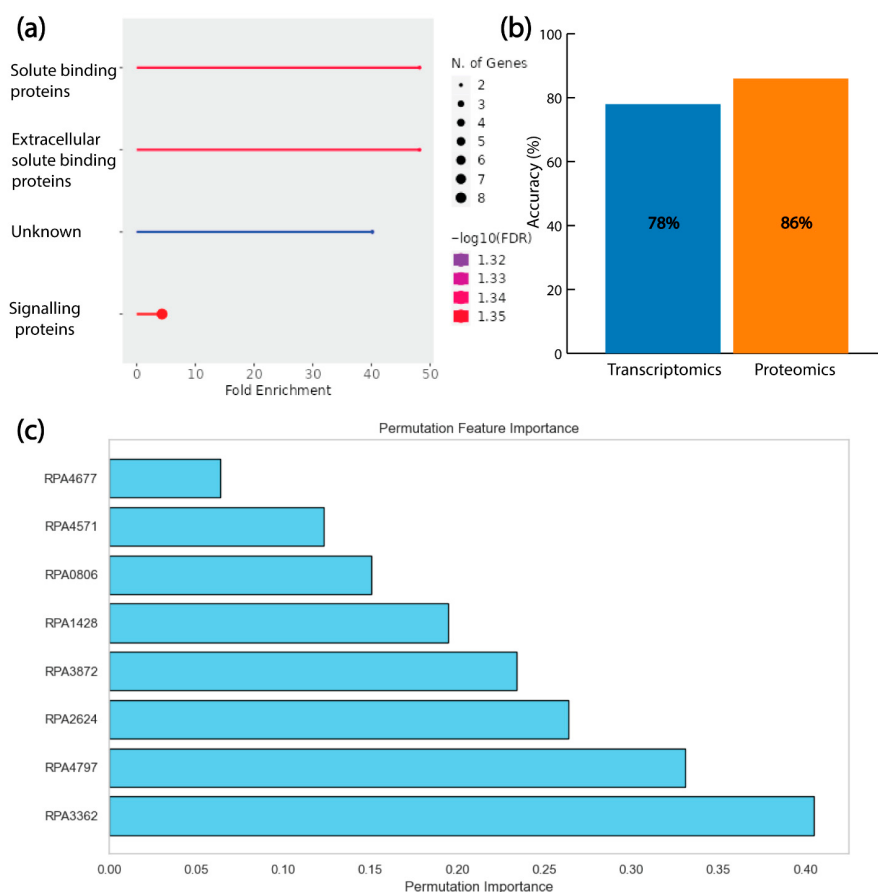


Figure 6. Only amino acid transport proteins can explain the growth pattern of *R. palustris* for LBPs. (a) Gene ontology analysis revealed the role of the transport protein as the signaling molecules. (b) Eight transport proteins can capture the growth rate in different conditions with 86% accuracy compared to the transcriptomics data's 78% accuracy. (c) Permutation feature importance revealed the top transport proteins.

4. Conclusions

This study generated growth profiles along with the transcriptomics and proteomics datasets for *R. palustris* during the catabolism of various lignin breakdown products (LBPs), including monolignols (*p*-coumaryl alcohol, coniferyl alcohol, sinapyl alcohol), acid derivatives (*p*-coumarate, sodium ferulate), and kraft lignin. These datasets are currently being used in a companion study to investigate common pathways for LBP metabolism in *R. palustris* [46]. Using transcriptomics and proteomics data as input features and growth rates as the target, three machine learning models—Artificial Neural Networks (ANNs), Random Forest (RF), and Support Vector Machines (SVM)—were trained, with the ANN model achieving the highest accuracy for both the transcriptomic (94%) and the proteomic (96%) datasets. Further analysis using permutation feature importance on the ANN models identified the top twenty genes and proteins influencing growth rates. By integrating the feature importance scores from both datasets, eight transport proteins were found to significantly impact *R. palustris* growth during LBP catabolism. Retraining the ANN model based on these eight transport proteins yielded prediction accuracies of 86% for proteomic data and 76% for transcriptomic data. Overall, this work demonstrates the utility of ANN models in identifying growth-associated genes and proteins that regulate the metabolic responses of *R. palustris* under aerobic and anaerobic conditions while catabolizing LBPs.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/metabo16010086/s1>, Figure S1: Growth curve of *R. palustris* in different LBPs for aerobic and anaerobic conditions; Figure S2: StringDB network of top twenty proteins; Figure S3: Co-efficient of variation for top nine differentially abundant between aerobic and anaerobic conditions; Table S1: Harvesting time for different conditions and growth rate calculations; Table S2: Top twenty genes impacting the growth rates; Table S3: Top twenty proteins impacting the growth rates; Table S4: Combined list of top twenty proteins.

Author Contributions: R.S. designed the study and oversaw the project and funding acquisition; N.B.C. performed data curation, formal analysis, validation, and visualization; N.B.C., M.K., and N.S. worked on methodology; N.B.C. wrote the original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a National Science Foundation (NSF) CAREER grant, grant number 1943310.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the codes required to reproduce the results of the manuscript can be accessed at https://github.com/ssbio/r_palustris_LBP_ML (accessed on: 11 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Shelton, J.L.; McIntosh, J.C.; Hunt, A.G.; Beebe, T.L.; Parker, A.D.; Warwick, P.D.; Drake, R.M.; McCray, J.E. Determining CO₂ storage potential during miscible CO₂ enhanced oil recovery: Noble gas and stable isotope tracers. *Int. J. Greenh. Gas Control* **2016**, *51*, 239–253. [\[CrossRef\]](#)
- Lei, X.; Lin, Y.; Yang, Q.; Zhou, J.; Chen, X.; Wen, J. Research on coordinated control of renewable-energy-based Heat-Power station system. *Appl. Energy* **2022**, *324*, 119736. [\[CrossRef\]](#)
- Furst, M.R.L.; Korkmaz, V.; Gaide, T.; Seidensticker, T.; Behr, A.; Vorholt, A.J. Tandem Reductive Hydroformylation of Castor Oil Derived Substrates and Catalyst Recycling by Selective Product Crystallization. *ChemCatChem* **2017**, *9*, 4319–4323. [\[CrossRef\]](#)
- Sun, X.; Atiyeh, H.K.; Li, M.; Chen, Y. Biochar facilitated bioprocessing and biorefinery for productions of biofuel and chemicals: A review. *Bioresour. Technol.* **2020**, *295*, 122252. [\[CrossRef\]](#)
- Li, C.; Zhao, X.; Wang, A.; Huber, G.W.; Zhang, T. Catalytic Transformation of Lignin for the Production of Chemicals and Fuels. *Chem. Rev.* **2015**, *115*, 11559–11624. [\[CrossRef\]](#)
- Ullah, M.; Liu, P.; Xie, S.; Sun, S. Recent Advancements and Challenges in Lignin Valorization: Green Routes towards Sustainable Bioproducts. *Molecules* **2022**, *27*, 6055. [\[CrossRef\]](#) [\[PubMed\]](#)
- VerBerkmoes, N.C.; Shah, M.B.; Lankford, P.K.; Pelletier, D.A.; Strader, M.B.; Tabb, D.L.; McDonald, W.H.; Barton, J.W.; Hurst, G.B.; Hauser, L.; et al. Determination and Comparison of the Baseline Proteomes of the Versatile Microbe *Rhodopseudomonas palustris* under Its Major Metabolic States. *J. Proteome Res.* **2006**, *5*, 287–298. [\[CrossRef\]](#) [\[PubMed\]](#)
- Larimer, F.W.; Chain, P.; Hauser, L.; Lamerdin, J.; Malfatti, S.; Do, L.; Land, M.L.; Pelletier, D.A.; Beatty, J.T.; Lang, A.S.; et al. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. *Nat. Biotechnol.* **2003**, *22*, 55–61. [\[CrossRef\]](#)
- Pan, C.; Oda, Y.; Lankford, P.K.; Zhang, B.; Samatova, N.F.; Pelletier, D.A.; Harwood, C.S.; Hettich, R.L. Characterization of Anaerobic Catabolism of p-Coumarate in *Rhodopseudomonas palustris* by Integrating Transcriptomics and Quantitative Proteomics. *Mol. Cell. Proteom.* **2008**, *7*, 938–948. [\[CrossRef\]](#)
- Orth, J.D.; Thiele, I.; Palsson, B.Ø. What is flux balance analysis? *Nat. Biotechnol.* **2010**, *28*, 245–248. [\[CrossRef\]](#)
- Alsiyabi, A.; Chowdhury, N.B.; Long, D.; Saha, R. Enhancing in silico strain design predictions through next generation metabolic modeling approaches. *Biotechnol. Adv.* **2022**, *54*, 107806. [\[CrossRef\]](#)
- Navid, A.; Jiao, Y.; Wong, S.E.; Pett-Ridge, J. System-level analysis of metabolic trade-offs during anaerobic photoheterotrophic growth in *Rhodopseudomonas palustris*. *BMC Bioinform.* **2019**, *20*, 233. [\[CrossRef\]](#)
- Alsiyabi, A.; Immethun, C.M.; Saha, R. Modeling the Interplay between Photosynthesis, CO₂ Fixation, and the Quinone Pool in a Purple Non-Sulfur Bacterium. *Sci. Rep.* **2019**, *9*, 12638. [\[CrossRef\]](#)

14. Chowdhury, N.B.; Alsiyabi, A.; Saha, R. Characterizing the Interplay of Rubisco and Nitrogenase Enzymes in Anaerobic-Photoheterotrophically Grown *Rhodopseudomonas palustris* CGA009 through a Genome-Scale Metabolic and Expression Model. *Microbiol. Spectr.* **2022**, *10*, e0146322. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Hartman, E.; Scott, A.M.; Karlsson, C.; Mohanty, T.; Vaara, S.T.; Linder, A.; Malmström, L.; Malmström, J. Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nat. Commun.* **2023**, *14*, 5359. [\[CrossRef\]](#)
16. Sundar, G.N.; Selvaraj, S.; Narmadha, D.; Sagayam, K.M.; Jone, A.A.A.; Aly, A.A.; Le, D.-N. An Intelligent Prediction Model for Target Protein Identification in Hepatic Carcinoma Using Novel Graph Theory and ANN Model. *Comput. Model. Eng. Sci.* **2022**, *133*, 31–46. [\[CrossRef\]](#)
17. Aytan-Aktug, D.; Clausen, P.T.L.C.; Bortolaia, V.; Aarestrup, F.M.; Lund, O. Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. *mSystems* **2020**, *5*, 10–1128. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Harwood, C.S.; Gibson, J. Anaerobic and aerobic metabolism of diverse aromatic compounds by the photosynthetic bacterium *Rhodopseudomonas palustris*. *Appl. Environ. Microbiol.* **1988**, *54*, 712–717. [\[CrossRef\]](#)
19. Oshlag, J.Z.; Ma, Y.; Morse, K.; Burger, B.T.; Lemke, R.A.; Karlen, S.D.; Myers, K.S.; Donohue, T.J.; Noguera, D.R. Anaerobic Degradation of Syringic Acid by an Adapted Strain of *Rhodopseudomonas palustris*. *Appl. Environ. Microbiol.* **2020**, *86*, e01888–19. [\[CrossRef\]](#) [\[PubMed\]](#)
20. O'Brien, E.J.; Lerman, J.A.; Chang, R.L.; Hyduke, D.R.; Palsson, B.Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **2013**, *9*, 693. [\[CrossRef\]](#)
21. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [\[CrossRef\]](#)
22. Ismail, W.; Gescher, J. Epoxy Coenzyme A Thioester Pathways for Degradation of Aromatic Compounds. *Appl. Environ. Microbiol.* **2012**, *78*, 5043–5051. [\[CrossRef\]](#)
23. Zaar, A.; Gescher, J.; Eisenreich, W.; Bacher, A.; Fuchs, G. New enzymes involved in aerobic benzoate metabolism in *Azoarcus evansii*. *Mol. Microbiol.* **2004**, *54*, 223–238. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Spero, M.A.; Aylward, F.O.; Currie, C.R.; Donohue, T.J. Phylogenomic Analysis and Predicted Physiological Role of the Proton-Translocating NADH:Quinone Oxidoreductase (Complex I) Across Bacteria. *mBio* **2015**, *6*, e00389–15. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Agnihotri, G.; Liu, H. Enoyl-CoA Hydratase: Reaction, Mechanism, and Inhibition. *ChemInform* **2003**, *34*, 9–20. [\[CrossRef\]](#)
26. Rahman, M.M.; Andberg, M.; Koivula, A.; Rouvinen, J.; Hakulinen, N. The crystal structure of D-xylonate dehydratase reveals functional features of enzymes from the Ilv/ED dehydratase family. *Sci. Rep.* **2018**, *8*, 865. [\[CrossRef\]](#)
27. Ge, S.X.; Jung, D.; Yao, R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **2020**, *36*, 2628–2629. [\[CrossRef\]](#)
28. Bozin, T.N.; Berdyshev, I.M.; Chukhontseva, K.N.; Karaseva, M.A.; Konarev, P.V.; Varizhuk, A.M.; Lesovoy, D.M.; Arseniev, A.S.; Kostrov, S.V.; Bocharov, E.V.; et al. NMR structure of emfourin, a novel protein metalloprotease inhibitor: Insights into the mechanism of action. *J. Biol. Chem.* **2023**, *299*, 104585. [\[CrossRef\]](#) [\[PubMed\]](#)
29. van der Ploeg, J.R.; Eichhorn, E.; Leisinger, T. Sulfonate-sulfur metabolism and its regulation in *Escherichia coli*. *Arch. Microbiol.* **2001**, *176*, 1–8. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Higgins, C.F. ABC transporters: Physiology, structure and mechanism—An overview. *Res. Microbiol.* **2001**, *152*, 205–210. [\[CrossRef\]](#)
31. Stuffle, E.C.; Johnson, M.S.; Watts, K.J. PAS domains in bacterial signal transduction. *Curr. Opin. Microbiol.* **2021**, *61*, 8–15. [\[CrossRef\]](#)
32. Chen, X.; Alakavuklar, M.A.; Fiebig, A.; Crosson, S. Cross-regulation in a three-component cell envelope stress signaling system of *Brucella*. *mBio* **2023**, *14*, e0238723. [\[CrossRef\]](#)
33. Bathke, J.; Konzer, A.; Remes, B.; McIntosh, M.; Klug, G. Comparative analyses of the variation of the transcriptome and proteome of *Rhodobacter sphaeroides* throughout growth. *BMC Genom.* **2019**, *20*, 358. [\[CrossRef\]](#)
34. Watkins-Dulaney, E.J.; Straathof, S.; Arnold, F.H. Tryptophan Synthase: Biocatalyst Extraordinaire. *ChemBioChem* **2020**, *22*, 5–16. [\[CrossRef\]](#)
35. Waters, N. Functional characterization of the acyl carrier protein (PfACP) and beta-ketoacyl ACP synthase III (PfKASIII) from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **2002**, *123*, 85–94. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Geissel, F.; Lang, L.; Husemann, B.; Morgan, B.; Deponte, M. Deciphering the mechanism of glutaredoxin-catalyzed roGFP2 redox sensing reveals a ternary complex with glutathione for protein disulfide reduction. *Nat. Commun.* **2024**, *15*, 1733. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Pechter, K.B.; Gallagher, L.; Pyles, H.; Manoil, C.S.; Harwood, C.S. Essential Genome of the Metabolically Versatile Alphaproteobacterium *Rhodopseudomonas palustris*. *J. Bacteriol.* **2016**, *198*, 867–876. [\[CrossRef\]](#)
38. Galperin, M.Y.; Wolf, Y.I.; Garushyants, S.K.; Alvarez, R.V.; Koonin, E.V. Nonessential Ribosomal Proteins in Bacteria and Archaea Identified Using Clusters of Orthologous Genes. *J. Bacteriol.* **2021**, *203*, 10–1128. [\[CrossRef\]](#) [\[PubMed\]](#)

39. Sun, G. Death and survival from executioner caspase activation. *Semin. Cell Dev. Biol.* **2023**, *156*, 66–73. [[CrossRef](#)]
40. Baker, M.D.; Neiditch, M.B. Structural Basis of Response Regulator Inhibition by a Bacterial Anti-Activator Protein. *PLoS Biol.* **2011**, *9*, e1001226. [[CrossRef](#)]
41. Shuai, H.; Myronovskyi, M.; Nadmid, S.; Luzhetskyy, A. Identification of a Biosynthetic Gene Cluster Responsible for the Production of a New Pyrrolopyrimidine Natural Product—Huimycin. *Biomolecules* **2020**, *10*, 1074. [[CrossRef](#)]
42. Nowrouzi, B.; Rios-Solis, L. Redox metabolism for improving whole-cell P450-catalysed terpenoid biosynthesis. *Crit. Rev. Biotechnol.* **2021**, *42*, 1213–1237. [[CrossRef](#)]
43. Fan, L.; Fan, L.; Yu, T.; Tan, X.; Shi, Z. Hydrothermal Synthesis of Lignin-Based Carbon Microspheres as Anode Material for Lithium-Ion Batteries. *Int. J. Electrochem. Sci.* **2020**, *15*, 1035–1043. [[CrossRef](#)]
44. Srinivasan, M.; Muthukumar, S.; Rajesh, D.; Kumar, V.; Rajakumar, R.; Akbarsha, M.A.; Gulyás, B.; Padmanabhan, P.; Archunan, G. The Exoproteome of *Staphylococcus pasteuri* Isolated from Cervical Mucus during the Estrus Phase in Water Buffalo (*Bubalus bubalis*). *Biomolecules* **2022**, *12*, 450. [[CrossRef](#)] [[PubMed](#)]
45. Kathol, M.; Chowdhury, N.B.; Immethun, C.; Alsiyabi, A.; Morris, D.; Naldrett, M.J.; Saha, R. High enzyme promiscuity in lignin degradation mechanisms in *Rhodopseudomonas palustris* CGA009. *Appl. Environ. Microbiol.* **2025**, *91*, e0057325. [[CrossRef](#)] [[PubMed](#)]
46. Immethun, C.M.; Kathol, M.; Changa, T.; Saha, R. Synthetic Biology Tool Development Advances Predictable Gene Expression in the Metabolically Versatile Soil Bacterium *Rhodopseudomonas palustris*. *Front. Bioeng. Biotechnol.* **2022**, *10*, 800734. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.