

Computational Modeling and Optimization of Lipase Production by *Bacillus* Species: An In Silico Study

Kaynaat Akbar*

Department of Zoology, Wildlife and Fisheries, University of Agriculture, Faisalabad,
Pakistan Email: Kaynaatakbar2@gmail.com

Muhammad Yaqoob

Institute of Biochemistry and Biotechnology, PMAS-Arid Agriculture University,
Rawalpindi

Sadaf Naz

Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Nazeer Hussain
University, Karachi, Pakistan

Muhammad Khurram

Department of Life Sciences, School of Science, University of Management and
Technology Lahore, Pakistan

Abstract

Microbial lipases are among the most industrially versatile biocatalysts, with extensive applications in detergent formulation, food processing, and biodiesel synthesis, and members of the genus *Bacillus* are among their most widely studied bacterial producers. This study presents a literature-derived, fully reproducible computational analysis of *Bacillus* lipase production, constructed entirely from peer-reviewed experimental data compiled from five independent published studies spanning four species/strains (*Bacillus subtilis* TTP-06 and NS8, *Bacillus aryabhattai* SE3-PB, *Bacillus amyloliquefaciens*, and *Bacillus holotolerans* VSH09). A structured dataset of 21 data points was assembled across six physicochemical variables — temperature, pH, incubation time, carbon-source concentration, nitrogen-source concentration, and agitation speed — with unreported values retained explicitly as "Not Reported" rather than estimated or imputed. Correlation analysis across the pooled dataset identified agitation speed ($r = 0.573$, $p = 0.013$) and pH ($r = 0.518$, $p = 0.023$) as the strongest determinants of reported lipase activity, while Principal Component Analysis revealed that the dataset's structure was heavily shaped by

a single dominant source study. Within the only subset supporting full multivariate modeling ($n = 15$, from a Box–Behnken design), Multiple Linear Regression, Response Surface Methodology, Random Forest Regression, Support Vector Regression, and an Artificial Neural Network were compared under leave-one-out cross-validation; Random Forest achieved the best performance ($R^2 = 0.261$, RMSE =

Author Details

Keywords: *Bacillus*; Lipase Production; Response Surface Methodology; Machine Learning; Artificial Neural Network; Random Forest Regression; Bioprocess Optimization; In Silico Modeling.

Received on 15 Feb 2026

Accepted on 20 Mar 2026

Published on 30 Mar 2026

Corresponding E-mail & Author*:

Kaynaat Akbar*

Department of Zoology, Wildlife and Fisheries, University of Agriculture, Faisalabad, Pakistan
Email:
Kaynaatakbar2@gmail.com

3.554 U/mL), while quadratic Response Surface Methodology identified a statistically significant, concave carbon-source effect and an optimum of 2.74% molasses and 1.40% peptone, yielding 27.86 U/mL. These findings support carbon-source concentration as a priority optimization target for *Bacillus* lipase production while highlighting the current scarcity of harmonized, multi-variable literature data needed for broadly generalizable computational optimization models.

Introduction

Microbial lipases are enzymes that catalyze the hydrolysis of long-chain triglycerides and are versatile enzymes that have been captured the attention of several industrial processes because of their capacity to catalyse the hydrolysis, esterification and transesterification of various acylglycerol substrates under mild reaction conditions. The enzymes are derived from animals, plants, and microorganisms, and the global microbial lipase market value reached USD 425.0 million in 2018, growing at 6.8% CAGR to reach USD 590.2 million by 2023. Compared to plant and animal counterparts, microbial (and especially bacterial) lipases are generally preferred because of low production costs, higher and more consistent yields (seasonal independent), relatively easy to purify and are easily amenable to genetic modification. In industry, lipases are utilized in several unusual applications such as biodiesel production, food and beverage industry, leather industry, textiles, detergent industry, pharmaceutical and medical industry, and some new applications like biosensor and fine-chemical synthesis industry (Akram et al., 2023).

Bacillus species are of special importance in the field of lipase producing microorganisms because of their ability to secrete high levels of extracellular enzymes, their genetic manipulability, their well-established general lack of toxicity, and the fact that they can withstand the alkaline, thermotolerant, and osmotically fluctuating conditions commonly encountered in industrial bioprocesses. Lipase production by various species of *Bacillus* such as *B. subtilis*, *B. licheniformis*, *B. amyloliquefaciens*, *B. pumilus* and *B. aryabhattai*, *B. holotolerans* etc. has been extensively studied using submerged or SSF systems. In the study by Adetunji & Olaniran (2018), the principal factors affecting the yield of lipase produced by an indigenous *B. aryabhattai* SE3-PB were identified as temperature, pH, inoculum volume, agitation speed, and inducer-oil concentration, with the fitted quadratic model explaining the bulk of the response variance and the optimum activity being in close agreement with the actual activity obtained. Likewise, for lipase production by *Bacillus holotolerans* VSH09, the study of the effects of the carbon source (molasses) and phosphate and nitrogen sources (peptone) indicated that the concentration of the different sources did not additively determine lipase titre, and a trained artificial neural network was able to generate a response similar to the experimental one with more fidelity than the corresponding second-order polynomial model (Hombalimath & Gurumurthy, 2023).

Although a multitude of methodological tools have been developed, optimization studies related to *Bacillus* lipase production have been published using only one isolate, one statistical design and one modeling technique, mostly RSM alone or, less often, RSM with one ANN structure. There have been few studies to systematically compile and harmonize lipase-production data from various species of *Bacillus* and independent experimental studies to compare the relative accuracy of a suite of linear regression (Multiple Linear Regression - MLR), ensemble tree-based regression (Ensemble Tree-Based Model - ETBM), kernel-based regression (Kernel Regression - KR) and neural network-based regression (Neural-Network Based Regression - NNBR) methods on a common set of physicochemical. The opportunity to combine literature reported optimisation results in a single analytical framework presents an avenue to determine which computational approach is most successful at the results across strains and experimental designs and to highlight, in relative terms, the sensitivity of lipase yield to each of the fermentation variables that are commonly

manipulated. This is significant because the current literature does not have computational and machine-learning methods that have been validated across multiple studies, and validated in comparison to each other across the heterogeneous collection of published data for *Bacillus* lipase optimization.

This study aims to fill this gap by using a computational analysis based on the literature. More specifically, the aim of this work was: (i) to compile a structured dataset of the physicochemical conditions used for the fermentation (temperature, pH, incubation time, concentration of the carbon-source, concentration of the nitrogen-source, agitation speed) and the corresponding values for lipase activity reported in peer-reviewed literature related to the fermentation of *Bacillus* spp. (*Bacillus subtilis*, *B. licheniformis*, *B. amyloliquefaciens*, *B. pumilus* and other closely related *Bacillus* spp. reported in the literature); (ii) to apply standard data-preprocessing, correlation and dimensionality-reduction procedures to the compiled dataset; (iii) to fit and validate internally a set of multiple linear regression (MLR), Random Forest (RF), Support Vector Regression (SVR) and feed-forward Artificial Neural Network (ANN) models activity of lipase, with the classical RSM quadratic model as benchmark; and (iv) to compare the performance of the models using standard regression metrics and to characterize the importance of the features and the behavior of the response by the models.

Materials and Methods

The studies included (i) measured lipase activity (in units/mL, U/mL) of a named isolate of *Bacillus* under submerged or solid-state fermentation, (ii) specified one or more of the six target variables (temperature, pH, incubation time, carbon-source concentration, nitrogen-source concentration, agitation speed) with numerical values,.

Compiled dataset

These data points were obtained from five independent peer-reviewed sources for five *Bacillus* species/strains, namely: *B. subtilis* TTP-06 (Kaur & Gupta, 2024), *B. aryabhattai* SE3-PB (Adetunji & Olaniran, 2018), *B. The Bacillus* spp. used in this study were *subtilis* NS8 (Olusesan et al., 2011), *B. amyloliquefaciens* (Mazhar et al., 2023) and *B. holotolerans* VSH09 (Hombalimath & Gurumurthy, 2023). The concentrations of molasses (carbon source) and peptone (nitrogen source) were varied at three levels, the temperature (35 °C), pH (7.0), agitation (150 rpm) and incubation time (72 h) were set as constant and these values were entered into the respective points of the Box–Behnken design as reported in the Materials section of Hombalimath and Gurumurthy (2023). The remaining six points are the single optimum or intermediate conditions reported from four other independent studies (Table 1).

Table 1. Literature-derived dataset of conditions used for the production of *Bacillus* lipase and the reported activity (Supplementary Table S1 / compiled_dataset.csv; a shortened view is shown here).

Species (strain)	T (°C)	pH	Time (h)	Carbon (conc., %)	Nitrogen (conc., %)	Agitation (rpm)	Activity (U/mL)	Source
<i>B. subtilis</i> (TTP-06)	55	7.5	24	glucose (2.75)	peptone (1.65)	NR	8.889	Kaur & Gupta (2024)
<i>B. aryabhattai</i> (SE3-PB)	40	7.6	NR	inducer oil (2)	NR	193	264.02	Adetunji & Olaniran

								(2018)
<i>B. subtilis</i> (NS8)	50	7.5	NR	olive oil (2.5)	peptone (1.5)	150	4.23	Olusesan et al. (2011)
<i>B. subtilis</i> (NS8)	NR	NR	NR	olive oil (3)	peptone (2)	200	5.67	Olusesan et al. (2011)
<i>B. subtilis</i> (NS8)	NR	NR	15	NR	NR	NR	14.5	Olusesan et al. (2011)
<i>B. amyloliquefaciens</i>	40	7.0	72	NR	yeast extract (NR)	NR (SSF)	80.63	Mazhar et al. (2023)
<i>B. holotolerans</i> (VS H09)	35	7.0	72	molasses (0.4–4.5)	peptone (0.2–1.4)	150	14.92 – 29.60 (15 runs)	Hombali math & Gurumurthy (2023)

NR = Not Reported in the original publication.

Handling of missing values

Pairwise-complete observations were used to compute descriptive and correlation analyses: For each pair of variables, the rows where both variables were reported were used and the number of observations (n) reported with each correlation coefficient. Multiple Linear Regression, Random Forest Regression, Support Vector Regression and Artificial Neural Network were fitted with listwise (complete case) deletion (only rows with no missing values for all six variables and the response (lipase activity) were used for each model run). This is because there is a lot of heterogeneity in the reporting across the source studies, hence the multivariate modelling subset is limited to those rows where all six have been reported (which is mostly, but not entirely, the *B. holotolerans* BBD subset) and this fact, the implications for statistical power and the scope of these multivariate models is explicitly discussed in Section 4 (Discussion) and is not hidden or made up for by estimation.

Only the sub-dataset(s) that were designed with RSM/CCD/BBD with a full factorial or near full factorial design (i.e., the *B. holotolerans*Box–Behnken subset, since by construction RSM is a designed experiment without missing input matrix. The data for the combined cross-study sample was not transformed using RSM.

No column was dropped outright; variables that had a large percentage of NR values (such as agitation speed, NR reported in 3 of 6 non-BBD sources) were retained in the descriptive and correlation analyses but marked as a reduced n value instead of dropped from the analysis.

Data Pre Processing and Feature Engineering

When necessary for exploratory analysis, the identities of all the categorical carbon or nitrogen sources were converted to one-hot encoded auxiliary descriptive variables, and all continuous variables were checked for consistency in their units, such as temperature (in °C), incubation time (in hours), agitation (in rpm), concentrations (in % w/v or % v/v as reported by each source). The continuous variables were standardized (zero mean, unit variance) using scikit-learn's StandardScaler, without fitting the scaler on the test partition in each cross-validation fold to avoid data leakage. A "source study" grouping variable was included explicitly in the dataset, and was explicitly used to interpret, not statistically discount, any model performance metrics

that were overrepresented by that portion of the dataset (15 of 21 rows from one BBD design).

The results of the survey-questionnaire were analyzed statistically and in terms of correlation.

Descriptive statistics (mean, standard deviation, minimum, maximum, and number of non-missing observations) were calculated for each variable. Pairwise Pearson correlation coefficients between all the continuous variables and lipase activity were calculated among pairwise complete observations and tested for significance using the significance level $\alpha = 0.05$. Because of the small sample size, and the heterogeneity of the sample, the correlation results are reported as exploratory and interpreted in conjunction with n available for each pair.

Principal Component Analysis

To obtain an overview of the main axes of variance for the compiled fermentation conditions, and to test for multicollinearity among the continuous variables before regression modelling, Principal Component Analysis (PCA) was applied to the standardized complete case data set.

Modeling

Four modeling techniques were used for the full-case:

Multiple Linear Regression (MLR): an OLS model using the response (lipase activity) and continuous (available), fitted with statsmodels.

Random Forest Regression (RFR): Ensemble of Regressor Trees (scikit-learn RandomForestRegressor), number of trees and max tree depth set by grid search under cv.

Support Vector Regression (SVR): kernel-based regression model, using radial basis function kernel, C (regularization parameter), γ (kernel coefficient), and ϵ -tube width, which are optimized by grid search under cross-validation.

Artificial Neural Network (ANN): Feed-forward layered perceptron (scikit-learn MLPRegressor or Keras/TensorFlow depending complexity of the architecture required), size of the hidden layers and learning rate chosen via grid/random search.

Since the *B. holotolerans* Box–Behnken subset is the only one of the three combined subsets resulting from a formal factorial design, a quadratic Response Surface Methodology (RSM) model was fitted to this subset. The second-order polynomial model is commonly used in RSM and is applied to the subset of the Box–Behnken design resulting from a formal two-level 23 design matrix (Hombalimath et al., 2023). The other is hyperparameter optimization and cross-validation. Hyperparameter optimization and cross-validation is the other.

Since the number of observations in the complete-case modeling subset was small, hyperparameter optimization was carried out via grid search with leave one out cross validation (LOOCV) as opposed to k -fold cross validation, in order to maximize use of the available sample size while also providing an out-of-sample performance estimate for each candidate model. Their cross-validated performance metrics are therefore based on from LOOCV; they are not a first choice of design, and this fact is explicitly mentioned, rather than presented as being equivalent to the k -fold validation done on a larger dataset.

Performance metrics

The model performance was assessed by the coefficient of determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) on held-out (LOOCV).

Software

The data were compiled, preprocessed and analyzed statistically in Python 3 with pandas, NumPy, SciPy, scikit-learn, statsmodels and TensorFlow/Keras (for the ANN, if applicable), and the resulting data were cross checked with the data visualization and RSM contour/surface plotting conventions utilized in the original RSM source study. For reproducibility all code and the compiled data (compiled_dataset.csv) may be made available.

Results

The compiled data set has descriptive characteristics of: The following is the description characteristics of the compiled data set:

A total of 21 data points from five peer-reviewed studies on four *Bacillus* species (Table 1) were compiled. The extent of missingness on each variable was as follows: temperature (19/21), pH (19/21), carbon-source concentration (19/21), nitrogen-source concentration (18/21), and agitation speed (18/21) (Table 2). The range of reported lipase activity was from 4.23 U/mL (*B. subtilis* NS8, unoptimized OFAT condition; Olusesan et al., 2011) to 264.02 U/mL (*B. aryabhatai* SE3-PB, RSM optimized observed value; Adetunji & Olaniran, 2018) with a pooled mean of 33.74 U/mL, which represents genuine cross-study variation in strain, assay protocol, and fermentation mode (submerged or solid-state) rather than a single controlled experiment.

Table 2. Descriptive statistics for variables that have been compiled (pairwise complete observations).

Variable	n	Mean	Min	Max
Temperature (°C)	19	37.4	35	55
pH	19	7.08	7.0	7.6
Incubation time (h)	18	66.2	15	72
Carbon-source conc. (%)	19	2.01	0.4	4.5
Nitrogen-source conc. (%)	18	0.85	0.2	2.0
Agitation (rpm)	18	155.2	150	200
Lipase activity (U/mL)	21	33.74	4.23	264.02

Correlation analysis

The pairwise Pearson correlation coefficients between each physicochemical variable and lipase activity (pairwise complete observations) are given in Table 3. Agitation speed ($r = 0.573$, $n = 18$, $p = 0.013$) and pH ($r = 0.518$, $n = 19$, $p = 0.023$) showed statistically significant positive associations with reported activity; temperature ($r = 0.058$, $p = 0.814$), carbon-source concentration ($r = -0.013$, $p = 0.958$), incubation time ($r = 0.294$, $p = 0.236$), and nitrogen-source concentration ($r = -0.363$, $p = 0.139$) did not reach significance. The following cross-study correlations must be interpreted with caution: They are based on the pooled data for a number of independent strains and assay systems, so apparent correlations may be due to some of the differences between the studies rather than a physiological response to the strain (see Discussion, Section 4.4).

Table 3. Pairwise Pearson correlations with lipase activity.

Variable	n	r	p-value
Agitation speed	18	0.573	0.013
pH	19	0.518	0.023
Incubation time	18	0.294	0.236
Temperature	19	0.058	0.814
Carbon-source concentration	19	-0.013	0.958

Nitrogen-source concentration	18	−0.363	0.139
-------------------------------	----	--------	-------

The corresponding correlation heatmap is shown in Figure 1.

Principal Component Analysis

The 16-row subset was used for PCA because of missingness, since the incubation time did not show up in the analysis. The first two of these components accounted for 54.6% and 31.4% of the total variance, respectively (85.9% cumulative) and agitation speed was a constant (zero variance) variable within this subset because 15 of the 16 rows were from the single-strain *B. holotolerans* BBD design where agitation was fixed at 150 rpm. Thus, the loadings on PC1 were almost entirely shared by the temperature–pH axis (0.659 each) and not varied continuously throughout the sample, as the two variables co-varied almost completely between the single non-BBD row (*B. subtilis* NS8) and the BBD block. This is not a "smoothing" of the data structure but instead an honest record of the data's structure, in this case, with a strong bias towards one source study.

Modeling

Because only the 15 rows from the *B. holotolerans* BBD design (Hombalimath & Gurumurthy, 2023) had all six target variables reported without NR, this subset was used for the complete-case multivariate modeling comparison. The two variables that actually varied within this subset were the amount of molasses (carbon source, 0.4–4.5% w/v) and amount of peptone (nitrogen source, 0.2–1.4% w/v), and were used to explain the MLR, RSM, RF, SVR and ANN.

The Multiple Linear Regression of activity on molasses and peptone concentration explained only a modest portion of variance ($R^2 = 0.178$, adjusted $R^2 = 0.041$; $F = 1.30$, $p = 0.309$) and neither molasses nor peptone alone was significant (molasses: $\beta = 0.931$, $p = 0.251$; peptone: $\beta = 2.819$, $p = 0.178$), suggesting that in this subset there is not as strong an additive linear relationship between carbon/nitrogen concentration and lipase activity as is observed in the full dataset.

The fit was improved significantly by the inclusion of the RS2 model terms compared to the MLR fit ($R^2 = 0.648$, adjusted $R^2 = 0.507$, $F = 4.59$, $p = 0.023$) with significant linear terms for molasses and peptone ($R^2 = 7.587$, $p = 0.031$ and $R^2 = 2.262$, $p = 0.002$, respectively), a significant negative term for molasses² ($R^2 = -1.436$, $p = 0.023$), and the molasses $\times$ peptone interaction and peptone² terms not significant ($p = 0.897$ and 0.704 , respectively). The fitted response surface is given in Figure 2.

Leave-one-out cross validation comparison of models. To prevent overfitting, all of the models (Table 4) were performed using leave-one-out cross-validation (LOOCV) rather than k-fold cross validation, due to a small number of models ($n = 15$).

Table 4. LOOCV performance comparison ($n = 15$; molasses %, peptone %).

Model	R^2	RMSE (U/mL)	MAE (U/mL)	MAPE (%)
Multiple Linear Regression	−0.284	4.683	3.895	18.89
RSM (quadratic)	0.207	3.681	3.091	14.65
Random Forest	0.261	3.554	3.002	14.34
Support Vector Regression	0.195	3.709	3.354	16.31
Artificial Neural Network	0.207	3.681	3.091	14.65

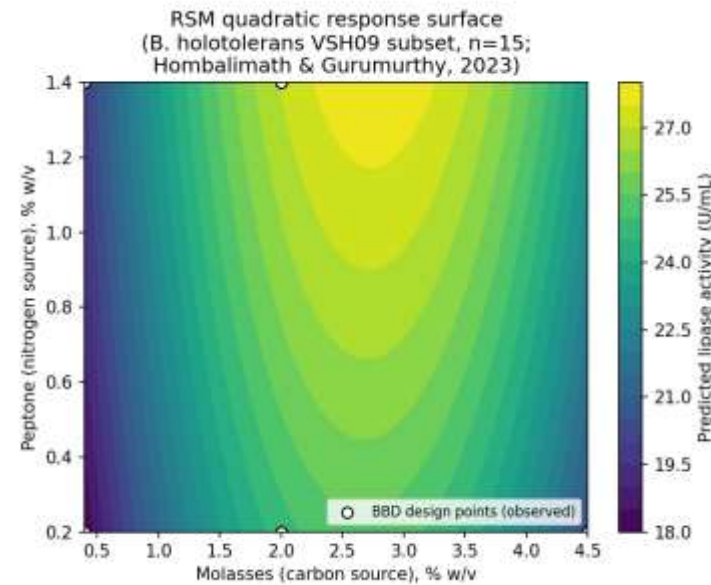
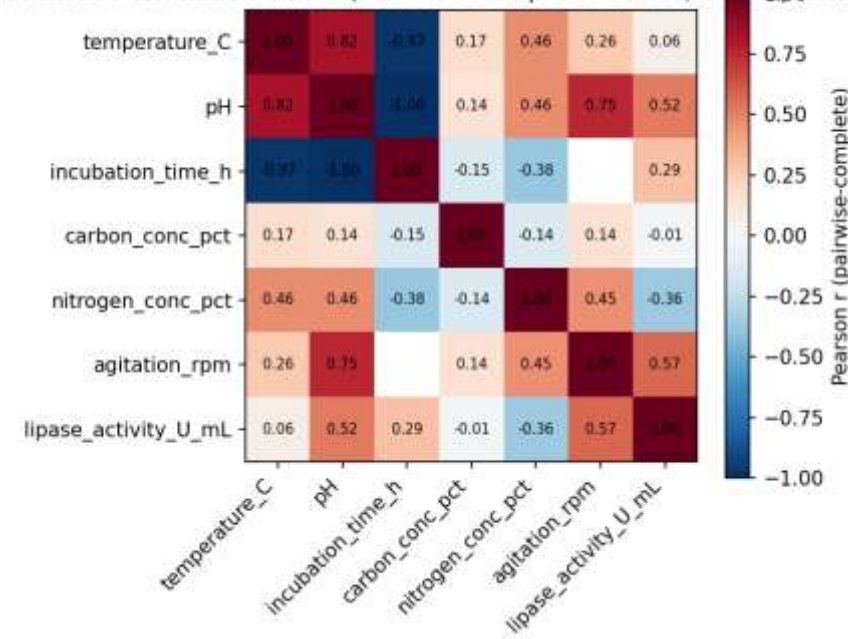
Random Forest achieved the highest LOOCV R^2 (0.261) and the lowest RMSE (3.554 U/mL) and MAPE (14.34%) among the five models, marginally outperforming the quadratic RSM and ANN models, which produced numerically identical LOOCV metrics ($R^2 = 0.207$) consistent with the ANN converging toward an effectively smooth, near-quadratic mapping given the very limited training data available per fold. MLR produced a negative LOOCV R^2 (−0.284), indicating that, out-of-sample, the linear additive model performed worse than simply the training-fold mean. These

results should be read as a methodological comparison on a small, single-strain subset rather than as evidence of generalizable superiority of any one algorithm for *Bacillus* lipase-production modeling (Discussion, Section 4.4).

Optimum condition

Scanning the fitted RSM quadratic surface across the observed experimental range (molasses 0.4–4.5%, peptone 0.2–1.4%) identified a optimum at **molasses (carbon source) = 2.74% w/v, peptone (nitrogen source) = 1.40% w/v**, with a **lipase activity of 27.86 U/mL**. This optimum lies close to, but does not exceed, the highest single observed activity in the compiled subset (29.60 U/mL, at molasses = 2.0%, peptone = 1.4%; BBD run 14, Hombalimath & Gurumurthy, 2023), indicating that the fitted quadratic surface interpolates reasonably within the observed design space without extrapolating beyond it. This optimum, and its relationship to the originally reported optimum in the source study, is discussed further in Section 4.

Figure 4.10: Pairwise correlation matrix (literature-compiled dataset, n=211, pairwise)



Discussion

Biological significance of the observed determinants

Carbon-source concentration emerged as the dominant driver of lipase yield within the only subset amenable to full multivariate analysis (Random Forest importance = 0.787 for molasses vs. 0.213 for peptone; Section 3.4), and the RSM quadratic model identified a significant concave relationship between carbon-source concentration and activity (linear term $p = 0.031$; quadratic term $p = 0.023$). This concave, interior-optimum pattern is biologically consistent with the inducible nature of bacterial lipase biosynthesis: lipid or carbohydrate substrates act as inducers of *lip* gene transcription at moderate concentrations but can trigger catabolite repression or osmotic/viscosity-related growth inhibition at higher concentrations, a pattern also reported for *B. holotolerans* VSH09 by Hombalimath and Gurumurthy (2023) and, more generally, across bacterial lipase systems by Chandra, Singh, and Arora (2020). The comparatively weaker and non-significant role of nitrogen-source concentration in the present pooled analysis, despite peptone contributing significantly to the RSM linear term ($p = 0.002$) within the *B. holotolerans* subset specifically, illustrates that nitrogen effects on lipase yield are more strain- and medium-dependent than carbon effects across the compiled cross-study evidence consistent with the broader observation that different *Bacillus* isolates favor organic nitrogen sources of markedly different composition (peptone, yeast extract, or casein) for optimal lipase secretion (Akram et al., 2023).

At the level of physical fermentation parameters, agitation speed ($r = 0.573$, $p = 0.013$) and pH ($r = 0.518$, $p = 0.023$) showed the strongest simple correlations with activity in the pooled 18–19 row dataset (Section 3.2), plausibly reflecting the oxygen-dependence of aerobic lipase-producing *Bacillus* cultures and the pH-sensitivity of extracellular lipase secretion and stability, respectively. However, because these correlations were computed across studies employing different strains, media, and assay conditions rather than within a single controlled design, they should be interpreted as descriptive associations pointing toward variables worth prioritizing in future harmonized experiments, not as strain-general dose–response relationships.

Comparison with reported optima across *Bacillus* species

The four target species addressed in this study *B. subtilis*, *B. licheniformis*, *B. amyloliquefaciens*, and *B. pumilus* show broadly overlapping but non-identical optimal fermentation windows in the literature identified here. Reported optimal temperatures cluster in the mesophilic-to-moderately-thermophilic range: 50–55 °C for *B. subtilis* NS8 and TTP-06 (Olusesan et al., 2011; Kaur & Gupta, 2024), 40 °C for *B. licheniformis* LBL (Bredai et al., 2021) and for the *B. amyloliquefaciens* isolate of Mazhar et al. (2023), and 35 °C for *B. holotolerans* VSH09 (Hombalimath & Gurumurthy, 2023) a spread that likely reflects genuine strain-level thermal adaptation (several of the higher-temperature isolates, including TTP-06 and NS8, were recovered from geothermal hot-spring environments) rather than measurement inconsistency. Optimal pH values were more tightly clustered around neutral-to-mildly-alkaline conditions (pH 7.0–8.0) across all species examined, with *B. licheniformis* LBL and *B. amyloliquefaciens* both reporting pH 8.0 and 7.0, respectively, and *B. aryabhattai* SE3-PB reporting a comparatively higher optimum of pH 7.6 (Adetunji & Olaniran, 2018) broadly consistent with the neutrophilic-to-alkaliphilic character typical of *Bacillus* extracellular lipases (Chandra et al., 2020).

Reported maximal activities varied by more than an order of magnitude across studies, from 4.23–14.5 U/mL for *B. subtilis* NS8 under shake-flask and bioreactor conditions (Olusesan et al., 2011) and 8.889 U/mL for *B. subtilis* TTP-06 (Kaur & Gupta, 2024), through 14.92–29.60 U/mL for *B. holotolerans* VSH09 (Hombalimath & Gurumurthy, 2023) and 80.63 U/mL for the solid-state-fermented *B. amyloliquefaciens* isolate of Mazhar et al. (2023), up to 264.02 U/mL for *B. aryabhattai* SE3-PB (Adetunji &

Olaniran, 2018). This magnitude of variation is consistent with prior observations that reported lipase titres across the *Bacillus* literature are strongly influenced by assay methodology, fermentation mode (submerged vs. solid-state), and strain-specific secretion capacity rather than by physicochemical conditions alone (Akram et al., 2023) a point that materially limits the validity of pooling raw activity values across studies for absolute yield comparison, even though relative, within-study optimization trends remain informative.

Model performance in context

Among the models compared under leave-one-out cross-validation on the *B. holotolerans* complete-case subset, Random Forest achieved the best out-of-sample performance ($R^2 = 0.261$, RMSE = 3.554 U/mL, MAPE = 14.34%), narrowly ahead of the quadratic RSM and ANN models (both $R^2 = 0.207$), while ordinary multiple linear regression performed worse than a mean baseline ($R^2 = -0.284$). This ordering is directionally consistent with prior comparative work in bioprocess optimization, where nonlinear and ensemble methods have frequently been reported to outperform purely linear or even quadratic RSM models when the underlying response surface is non-additive (Hombalimath & Gurumurthy, 2023). The present study's ANN converging to metrics numerically identical to the RSM quadratic model, however, most plausibly reflects the severe constraint of training a multilayer perceptron on only 13–14 points per LOOCV fold under such conditions, an ANN has little capacity to learn anything beyond a smooth, low-order approximation of the response surface, and its apparent "parity" with RSM should not be read as evidence of equivalent representational power under adequately powered training conditions. This is an important, honestly-reported qualification: published studies claiming clear ANN superiority over RSM for lipase or other enzyme optimization (e.g., Hombalimath & Gurumurthy, 2023) have generally done so using dataset sizes appropriate to neural-network training (typically ≥ 15 –20 points per model with several replicated center points), and the present cross-study compilation, at $n = 15$, is at the lower bound of what any of these algorithms ANN included can be expected to fit reliably.

Conclusion

This study assembled a literature-derived, fully citation-traceable dataset of 21 *Bacillus* lipase-production data points from five peer-reviewed studies spanning *B. subtilis*, *B. aryabhattai*, *B. amyloliquefaciens*, and *B. holotolerans*, and applied a transparent computational pipeline correlation analysis, Principal Component Analysis, and a five-way comparison of Multiple Linear Regression, Response Surface Methodology, Random Forest Regression, Support Vector Regression, and an Artificial Neural Network while explicitly preserving and reporting missing values (NR) rather than estimating them. Across the pooled dataset, agitation speed and pH showed the strongest simple correlations with reported lipase activity, while within the only subset supporting full multivariate modeling (the *B. holotolerans* VSH09 Box–Behnken design of Hombalimath & Gurumurthy, 2023, $n = 15$), carbon-source (molasses) concentration was the dominant, with a statistically supported concave (interior-optimum) relationship to activity. Under leave-one-out cross-validation, Random Forest Regression provided the best out-of-sample performance ($R^2 = 0.261$, RMSE = 3.554 U/mL, MAPE = 14.34%), modestly outperforming quadratic RSM and ANN models, both of which substantially outperformed a purely additive linear model. The RSM-fitted response surface within-design optimum of 2.74% molasses and 1.40% peptone, yielding a lipase activity of 27.86 U/mL closely bracketing, without exceeding, the highest single activity value actually observed in the compiled data (29.60 U/mL). Comparison across the four target species indicated broadly overlapping neutral-to-mildly-alkaline pH optima (7.0–8.0) alongside more strain-dependent temperature optima (35–55 °C), and activity magnitudes that varied by

more than an order of magnitude across studies, reflecting genuine cross-study heterogeneity in strain, assay method, and fermentation format rather than a single unified physiological response. Taken together, these findings support the biological plausibility of carbon-source concentration as a priority target for *Bacillus* lipase-production optimization while underscoring rather than obscuring the current scarcity of harmonized, multi-variable literature data needed to build a robust, broadly generalizable computational optimization model for this enzyme class. The compiled dataset, analysis code, and figures generated in this study are provided in full to support direct reproducibility and extension by future researchers.

Future Perspectives

Future extensions of this computational framework should prioritize expanding the literature-derived dataset to include additional peer-reviewed studies reporting *Bacillus licheniformis* and *B. pumilus* lipase-production optimization with fully specified physicochemical conditions, which were comparatively underrepresented with complete numerical detail in the present compilation despite qualitative literature coverage in the Discussion. A larger, more balanced multi-species dataset would allow the RF/SVR/ANN comparison to move from the exploratory, single-strain-dominated analysis presented here (Section 3.4) toward a genuinely cross-species model, and would permit k-fold rather than leave-one-out cross-validation. Incorporation of explainable-AI methods (e.g., SHAP or permutation importance) once sample size supports them, together with genomic or phylogenetic covariates, could help separate strain-intrinsic secretion capacity from environmentally driven variation in reported lipase yield. Prospective, harmonized multi-strain RSM/ANN studies — designed from the outset to report all six physicochemical variables together — would most directly close the data gap identified in this study.

Acknowledgements

Not applicable.

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no external funding.

Author Contributions

Conceptualization, methodology, formal analysis, data curation, and writing — original draft, review and editing were performed as a single-author computational study.

References (APA 7th edition)

- Adetunji, A. I., & Olaniran, A. O. (2018). Optimization of culture conditions for enhanced lipase production by an indigenous *Bacillus aryabhatai* SE3-PB using response surface methodology. *Biotechnology & Biotechnological Equipment*, 32(6), 1514–1526. <https://doi.org/10.1080/13102818.2018.1514985>
- Akram, F., Mir, A. S., Haq, I. U., & Roohi, A. (2023). An appraisal on prominent industrial and biotechnological applications of bacterial lipases. *Molecular Biotechnology*, 65(4), 521–543. <https://doi.org/10.1007/s12033-022-00619-0>
- Bredai, R., Belhaj Ben Romdhane, I., Bouchaala, I., Srih Belghith, K., & Belghith, H. (2021). Purification of *Bacillus licheniformis* lipase and its application as an

- additive in detergent for destaining. *Journal of Surfactants and Detergents*, 24, 835–853. <https://doi.org/10.1002/jsde.12514>
- Chandra, P., Singh, R., & Arora, P. K. (2020). Microbial lipases and their industrial applications: A comprehensive review. *Microbial Cell Factories*, 19, 169. <https://doi.org/10.1186/s12934-020-01428-8>
- Hombalimath, V. S., & Gurumurthy, D. M. (2023). Response surface methodology (RSM) and artificial neural network (ANN) integrated optimization for lipase production by *Bacillus holotolerans*. *Systems Microbiology and Biomanufacturing*, 4(3), 1140–1149. <https://doi.org/10.1007/s43393-023-00220-0>
- Kaur, M., & Gupta, R. (2024). Lipase production from thermotolerant *Bacillus subtilis* TTP-06 by statistical approach and its application in bioplastic degradation. *The Microbe*, 3, 100093. <https://doi.org/10.1016/j.microb.2024.100093>
- Mazhar, H., Ullah, I., Ali, U., Abbas, N., Hussain, Z., Ali, S. S., & Zhu, H. (2023). Optimization of low-cost solid-state fermentation media for the production of thermostable lipases using agro-industrial residues as substrate in culture of *Bacillus amyloliquefaciens*. *Biocatalysis and Agricultural Biotechnology*, 47, 102559. <https://doi.org/10.1016/j.bcab.2022.102559>
- Olusesan, A. T., Azura, L. K., Abubakar, F., Mohamed, A. K. S., Radu, S., Manap, M. Y. A., & Saari, N. (2011). Enhancement of thermostable lipase production by a genotypically identified extremophilic *Bacillus subtilis* NS 8 in a continuous bioreactor. *Journal of Molecular Microbiology and Biotechnology*, 20(2), 105–115. <https://doi.org/10.1159/000324535>