processes

MDPI

Article

Near-Infrared Spectroscopy for Online Glucose Detection in Fermentation Processes: Transflectance/Transmission Sensor Evaluation and Modeling Optimization

Sipeng Yang [1,2], Zhikai Liu [3], Junbing Tao [1,2], Fengxu Xiao [1,2], Guiyang Shi [1,2] and Youran Li [1,2,*]

[1] Key Laboratory of Carbohydrate Chemistry and Biotechnology of Ministry of Education, School of Biotechnology, Jiangnan University, Wuxi 214122, China; 6230209077@stu.jiangnan.edu.cn (S.Y.); 1022220127@stu.jiangnan.edu.cn (J.T.); 8202306022@jiangnan.edu.cn (F.X.); gyshi@jiangnan.edu.cn (G.S.)
[2] National Engineering Research Center of Cereal Fermentation and Food Biomanufacturing, Jiangnan University, Wuxi 214122, China
[3] Wuxi Institute for Specialized Nutrition and Health Co., Ltd., Wuxi 214122, China; zhikai.liu@teshiyuan.com
* Correspondence: liyouran@jiangnan.edu.cn

Abstract

This study employed near-infrared (NIR) spectroscopy for real-time spectral acquisition of fermentation broth in lab-scale bioreactors, comparing the performance of transflectance and transmission sensors through glucose modeling and prediction while optimizing modeling approaches. The results demonstrated superior adaptability of transflectance sensors in fermentation environments: in conventional fermentation, glucose models exhibited lower errors (RMSEC = 4.087 g/L, RMSEV = 9.829 g/L) compared to transmission sensors (RMSEC = 5.972 g/L, RMSEV = 10.904 g/L), with significantly higher predictive performance (RPD = 3.735 vs. 2.369), indicating enhanced fitting accuracy and stability. In complex natural media containing peptone and yeast extract, transmission sensor performance deteriorated dramatically due to turbidity interference (R^2_{cal} = 0.134), whereas transflectance sensors maintained robust performance (R^2_{cal} = 0.993), confirming their adaptability to complex matrices. Regarding modeling strategies, the 1550–1700 nm spectral region demonstrated optimal feature extraction capability (RMSEC = 3.269 g/L, R^2_{cal} = 0.987). Basic preprocessing methods such as the moving average smoothing method have become the preferred preprocessing methods, as they strike a balance between calibration and prediction performance. Outlier removal analysis revealed that moderate elimination of 12 high-error samples (accounting for 30% of the total 39 samples) reduced RMSEC to 1.441 g/L and improved R^2_{cv} to 0.996, optimizing model performance; however, excessive removal of outlier samples degraded model capability, necessitating judicious sample selection. For fixed total sample sizes, calibration sets comprising 70–80% of samples yielded more reliable predictions. In conclusion, transflectance sensors demonstrate superior compatibility with multicomponent fermentation systems. Combined with wavelength selection, moving average preprocessing, and rational sample removal and partitioning strategies, this approach provides an effective solution for NIR-based online glucose monitoring.

Keywords: near-infrared spectroscopy; fermentation process; online monitoring; partial least squares regression

Check for updates

Academic Editor: Pedro Fernandes

Received: 14 February 2026
Revised: 9 March 2026
Accepted: 17 March 2026
Published: 25 March 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) license.

1. Introduction

Modern industrial fermentation is a sophisticated biomanufacturing paradigm for producing high-value chemicals and pharmaceuticals [1,2], requiring precise metabolic control to optimize yield and productivity [3–5]. Among critical parameters, glucose concentration is the most pivotal, as it regulates both microbial energy and global gene expression. Deviations from the optimal physiological window are costly: excess glucose triggers detrimental overflow metabolism (e.g., acetate formation or the Crabtree effect), while starvation induces stress and protein degradation. Despite its importance, real-time chemical monitoring remains a significant bottleneck in Process Analytical Technology (PAT) compared to routine physical parameters [6]. Traditionally, glucose analysis relies heavily on offline sampling followed by laboratory assays, such as High-Performance Liquid Chromatography (HPLC) or enzymatic biochemical analyzers. Although HPLC offers high accuracy and specificity, the workflow is inherently discontinuous, labor-intensive, and suffers from a significant time delay between sampling and data acquisition—often ranging from 30 min to several hours. This latency renders "real-time" feedback control impossible, relegating operators to reactive rather than proactive process management [7]. While online enzymatic biosensors and auto-analyzers have been developed to bridge this gap, they frequently suffer from limited operational stability; the bioactive enzymes immobilized on these sensors are susceptible to fouling by proteins and cells in the fermentation broth and often cannot withstand the harsh high-temperature sterilization (autoclaving) procedures required for bioreactors. Thus, there is an urgent industry-wide demand for robust, non-invasive, and rapid PAT tools capable of providing continuous insight into the fermentation broth's chemical composition.

Near-infrared (NIR) spectroscopy has emerged as a frontrunner among potential PAT tools for bioprocess monitoring due to its non-destructive nature, minimal sample preparation requirements, and rapid acquisition speed. Spanning the electromagnetic spectrum from 780 to 2500 nm [8,9], NIR spectroscopy probes the vibrational transitions of chemical bonds, specifically the overtones and combination bands of hydrogen-containing functional groups (C–H, O–H, N–H) [10,11]. Since fermentation broths are chemically complex mixtures dominated by water, carbohydrates, proteins, and cellular biomass, they exhibit rich, albeit complex, spectral signatures in the NIR region. When NIR radiation interacts with a fermentation sample, specific wavelengths are absorbed by analyte molecules according to their chemical structure. By analyzing the attenuation of light intensity and applying advanced chemometric algorithms—such as Partial Least Squares (PLS) regression—it is possible to deconvolute these overlapping spectral signals and quantitatively predict multiple parameters simultaneously. Consequently, NIR spectroscopy has found extensive applications across diverse fields [12–15], particularly in the food industry [16–19], where it enables rapid compositional analysis of feed ingredients in animal nutrition [20–22] and facilitates quality monitoring and screening of pharmaceutical tablets [23]. Of course, NIR spectroscopic instruments vary considerably in performance characteristics, with different spectrometer types exhibiting disparities in resolution and wavelength accuracy [24,25]. For instance, Fourier-transform spectrometers achieve resolutions of approximately 16 nm, whereas filter-based instruments attain 0.5 nm resolution [26]. Second, the stability of constructed calibration models remains constrained by multiple factors, including environmental fluctuations, instrument aging, and variations in sample physical properties [27]. Third, NIR detection demonstrates insufficient sensitivity for low-concentration analytes [28], primarily due to its absorption mechanism, which typically requires analyte concentrations exceeding 0.1%.

Translating NIR spectroscopy from solid-state analysis to the dynamic environment of a bioreactor remains challenging due to the aqueous matrix's spectral dominance

and significant light scattering within turbid fermentation broths. The strong absorption bands of water frequently mask weaker analyte signals, while multiple scattering events (Mie scattering)—driven by high biomass concentrations and complex industrial substrates—induce non-linear deviations and baseline shifts by rendering the effective optical path length undefined. Consequently, there remains a critical need for robust chemometric models capable of navigating the optical complexity of naturally derived media, which often exceeds the simplicity of the synthetic media used in the current literature.

The selection of an appropriate optical interface is critical for robust online monitoring in bioprocesses. Transmission probes operate on a direct alignment of the light source and detector across a fixed gap, governed by the Beer–Lambert Law. While effective for clarified liquids, they suffer from extreme signal attenuation and low signal-to-noise ratios in high-turbidity fermentation broths where biomass induces significant photon scattering [12]. In contrast, transflectance probes utilize an integrated reflector to capture both transmitted and back-scattered light, effectively doubling the optical path and enhancing signal collection efficiency in dense scattering media [29]. Despite these theoretical distinctions, empirical side-by-side comparisons of these geometries in real-time fermentation remain scarce. This study evaluates their adaptability in high-cell-density environments, employing a comprehensive chemometric strategy—including rigorous preprocessing and characteristic wavelength selection (e.g., glucose C-H bands)—to mitigate hardware-induced noise and ensure long-term monitoring stability.

In this study, we established an integrated experimental platform coupling a 5 L bioreactor with an external circulation loop containing both transmission and transflectance flow cells in series. This unique setup allows for the simultaneous acquisition of spectral data from both sensor types on the exact same fermentation broth, strictly eliminating sampling variability. By subjecting both sensors to identical biological and chemical fluctuations in real-time, this approach provides a rigorous, high-fidelity benchmark that overcomes the limitations of separate trials often found in the literature. The primary objectives of this work are to systematically evaluate and compare the performance of transflectance versus transmission sensors in online glucose monitoring, specifically investigating their robustness when switching from a defined synthetic medium to a high-turbidity complex medium containing peptone and yeast extract; to quantify the impact of matrix complexity on the signal-to-noise ratio and prediction accuracy; and to develop a comprehensive modeling framework that optimizes wavelength selection, preprocessing techniques, and rational outlier removal strategies. By integrating hardware comparison with advanced chemometric optimization, this study aims to provide a validated, robust solution for online glucose monitoring in complex fermentation systems.

2. Materials and Methods

2.1. Materials

Strain and Culture Media

The *Bacillus licheniformis* strain used in this study was provided by the School of Biotechnology and Key Laboratory of Carbohydrate Chemistry and Biotechnology of Ministry of Education, Jiangnan University. This engineered strain is a kanamycin-resistant derivative with the corresponding resistance trait. *Bacillus licheniformis* was selected for this study because of its excellent protein secretion capacity, well-characterized genetic background, and its status as a 'Generally Recognized as Safe' (GRAS) industrial workhorse.

The peptone, yeast extract, and agar powder for culturing related reagents were all purchased from OXOID Company in the UK. The antibiotic kanamycin and other reagents used in the experiment were all purchased from Aladdin Reagents Co., Ltd., Shanghai, China.

LB broth was used as the activation medium for seed culture. Composition of LB broth: NaCl 10 g/L, peptone 10 g/L, yeast extract 5 g/L.

Synthetic medium and complex medium were used as fermentation media.

Synthetic medium: K_2HPO_4 10 g/L, KH_2PO_4 1.36 g/L, $(NH_4)_2SO_4$ 5 g/L, urea 10 g/L, $MgSO_4 \cdot 7H_2O$ 2 g/L, monosodium glutamate 10 g/L, glucose 30 g/L, trace element solution (MKF) 2 g/L.

Complex medium: K_2HPO_4 10 g/L, KH_2PO_4 1.36 g/L, $(NH_4)_2SO_4$ 5 g/L, urea 10 g/L, $MgSO_4 \cdot 7H_2O$ 2 g/L, monosodium glutamate 10 g/L, glucose 30 g/L, MKF (trace elements) 2 g/L, peptone 20 g/L, yeast extract 10 g/L. The specific composition and concentration of the trace element solution MKF (micronutrients) are shown in Table 1.

Table 1. Composition and Concentration of Trace Element Solution (MKF).

MKF (Trace Element Solution) Component	Concentration (g/L)
Ferric ammonium citrate	54.4
$MnCl_2 \cdot 4H_2O$	9.8
$CoCl_2 \cdot 6H_2O$	1.6
$CuCl_2 \cdot 2H_2O$	1
H_3BO_3 (boric acid)	1.9
$ZnSO_4 \cdot 7H_2O$	9
$Na_2MoO_4 \cdot 2H_2O$	1.1
Na_2SeO_3	1.5
$NiSO_4 \cdot 6H_2O$	1.5

2.2. Fermentation Procedures

2.2.1. Seed Culture Preparation

The *Bacillus licheniformis* strain preserved in glycerol stocks at −70 °C was streaked onto agar plates supplemented with kanamycin sulfate (final concentration 50 μg/mL) and incubated at 37 °C for a minimum of 12 h.

Primary seed culture: A single colony from the cultured plate was inoculated into a 250 mL Erlenmeyer flask containing 15 mL LB medium supplemented with kanamycin sulfate (50 μg/mL final concentration) and cultivated at 37 °C with orbital shaking at 220 rpm for 18 h.

Secondary seed culture: The primary seed culture was transferred at a 9% (*v*/*v*) inoculation ratio into a flask containing 30 mL LB medium with kanamycin sulfate (50 μg/mL final concentration) and incubated under identical conditions (37 °C, 220 rpm) for 18 h.

2.2.2. Fermentation Cultivation

A 5 L glass fermenter was employed with an initial working volume of 2 L, an aeration rate of 1 vvm (aerobic cultivation), and a temperature maintained at 37 °C. Two fermentation batches were conducted using synthetic and natural media, respectively. To avoid excessive shear stress inhibiting cell growth at the early stage, the initial agitation speed was set at 500 rpm; as cell density and oxygen consumption increased during the middle stage, the agitation speed was manually adjusted to 800 rpm to improve dissolved oxygen supply. The pH was controlled within the range of 6.8–7.2 using 6 mol/L NaOH.

2.3. Construction of the NIR Experimental Platform for the 5 L Fermenter

This experiment utilizes a self-built experimental platform, as shown in Figure 1. The transmission and trans-reflection flow cells are connected in series, allowing the fermentation broth to sequentially pass through both cells. The flow cells are connected to a near-infrared spectrometer, enabling real-time spectral data acquisition of the fermentation broth. The two spectrometers used in this study are of the transflectance and transmission

types, respectively. To ensure the successful completion of the instrument self-check, the integration times of the transflectance and transmission spectrometers are set at 80 ms and 100 ms, respectively. Given that the fermentation process is carried out continuously, the acquisition interval of the two instruments is set to 60 s to achieve continuous data collection, with both having an averaging number of 60 and a subsample number of 1. This means that each recorded spectrum was obtained by averaging 60 consecutive scans, which is a standard acquisition setting of the spectrometers. (The near-infrared spectrometer was produced by Wuxi Xunjie Guangyuan Technology Co., Ltd., Wuxi, China, and the 5 L fermenter was produced by Shanghai Biotech Co., Ltd., Shanghai, China).

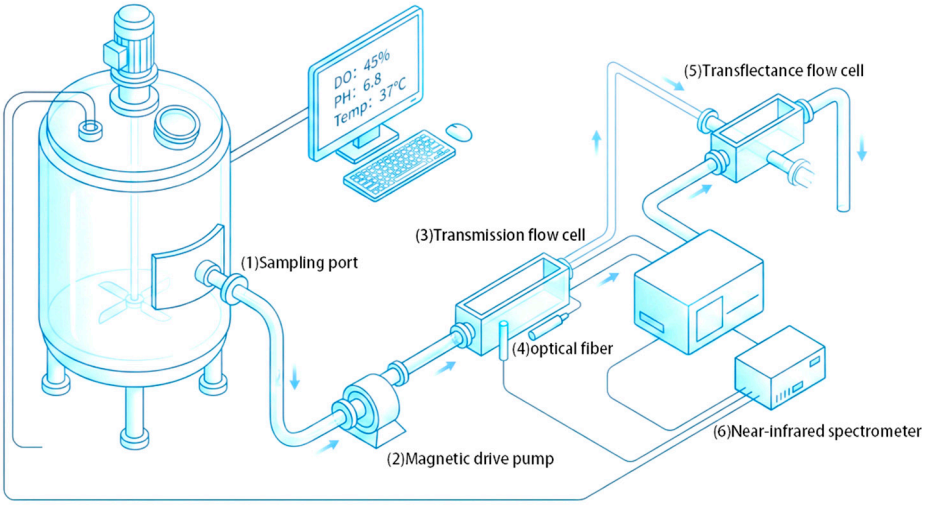


Figure 1. Schematic diagram of the fermentation tank and near-infrared acquisition system connection.

The transmission sensor operates by installing two optical fibers at opposite ends of the flow cell. Light emitted from one end traverses the fermentation broth and reaches the other end, where spectral information is captured by the instrument. In contrast, the transflectance sensor utilizes an insertion probe positioned within the flow cell. This probe contains an integrated optical fiber with emission and reception capabilities at its hollow tip. Light penetrates the fermentation broth to reach the receiving end, where a reflective device redirects the light back to the emission point, thereby capturing both transmission and reflection information.

Following inoculation, fermentation commenced with the magnetic pump flow rate set at 4 L/min. The fermentation broth circulated through a three-way valve to the magnetic pump, then bifurcated to the transmission and transflectance flow cells before returning to the fermenter, completing a full circulation cycle. The transmission mode employed fiber optic probes positioned on opposite sides of the flow cell, while the transflectance mode utilized a probe-based configuration within the flow cell.

2.4. Analytical Methods

Determination of Glucose Concentration

A 1 mL fermentation broth sample was centrifuged at 12,000 rpm for 20 min. Subsequently, 500 µL of the supernatant was mixed with 500 µL of trichloroacetic acid (10%) and incubated at 4 °C for 4 h to precipitate proteins. After centrifugation at 12,000 rpm for 20 min, the supernatant was filtered through a 0.22 µm aqueous membrane using a syringe and transferred to HPLC vials. Analysis was performed using a routine laboratory-established HPLC method for glucose determination. Separation was carried out with ultrapure water as the mobile phase, a Dikma CarboPac Ca^{2+} column (300 mm × 8.0 mm,

6 μm), a flow rate of 0.8 mL/min, and a column temperature of 80 °C. Detection was achieved using a refractive index (RI) detector maintained at 35 °C.

2.5. Model Development

2.5.1. Modeling Software

The IAS Cal-Box standalone modeling software from Wuxi Xunjie Guangyuan was utilized.

2.5.2. Model Performance Evaluation Metrics

Model performance was primarily assessed through key statistical indicators, including the root mean square error of calibration (RMSEC), root mean square error of prediction (RMSEP), and root mean square error of cross-validation (RMSECV). The root mean square error reflects the average prediction error, with lower values indicating higher model accuracy. The coefficient of determination (R^2) represents the proportion of variance in the dependent variable explained by the model, ranging from 0 to 1, where values closer to 1 indicate superior model fit [30,31].

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-1}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \quad (2)$$

where n denotes the number of samples; y_i represents the actual measured value of the ith sample; $\hat{y}_i$ represents the model-predicted value of the ith sample; and $\bar{y}$ represents the mean of the actual values.

The ratio of performance to deviation (RPD) is another commonly employed evaluation metric [32], calculated as follows:

$$\mathrm{RPD} = \mathrm{SD}/\mathrm{RMSEP} \quad (3)$$

The corresponding criterion is as follows: RPD > 2 indicates that the model possesses high reliability.

The Range Error Ratio (RER) was employed to assess the model's practical utility across the analytical range. It is defined as the ratio of the range of the validation set to the RMSEP:

$$\mathrm{RER} = \frac{y_{max} - y_{min}}{\mathrm{RMSEP}} \quad (4)$$

where y_{max} and y_{min} are the maximum and minimum reference values in the validation set, respectively. In this regard, values of less than 6 are considered poor; those from 7 to 10 represent a moderate level suitable for screening; those from 10 to 15 denote a good level applicable to quality control; and values greater than 20 correspond to an excellent level for high-precision quantitative analysis.

2.5.3. PLS Algorithm Modeling

The acquired NIR spectral data were correlated with actual measured values (reference values), and a quantitative analysis model was established using the partial least squares (PLS) algorithm. Internal validation of the quantitative model was subsequently performed using leave-one-out cross-validation [33,34]. This study adopts this validation method because the sample size is not large. The aim is to make the most of the limited data, reduce

the randomness of the validation results, and ensure that, under strict univariate controls, the model results of multiple independent sub-experiments are consistent and comparable.

The PLS model was built with leave-one-out cross-validation (LOOCV) (software default). The optimal LVs were automatically selected by the software based on the minimum RMSECV, a standard criterion for PLS optimization in NIR analysis, to ensure prediction accuracy and avoid overfitting.

2.6. Experimental Design

The experimental design of this study was structured into two core research sections, namely the adaptability and comparison of sensors under fermentation scenarios, and the systematic investigation of chemometric modeling strategies for fermentation process analysis. The detailed design and implementation principles of each section are elaborated as follows.

2.6.1. Sensor Adaptability and Comparison in Fermentation Scenarios

This section was dedicated to evaluating the applicability of sensors for on-line detection in fermentation processes, and the performance comparison was carried out from two complementary dimensions. First, the adaptability of candidate sensors was compared and verified under a single fermentation scenario to screen out sensors with basic detection feasibility for fermentation systems. Second, the adaptability and detection performance of the two sensors were further compared across different fermentation scenarios constructed by varying fermentation media. This dual-dimensional comparison was designed to comprehensively assess the compatibility of sensors with the complex and variable microenvironments of fermentation processes, and to determine the optimal sensor prototype for subsequent on-line modeling, thus laying a reliable hardware foundation for the establishment of fermentation process prediction models.

2.6.2. Optimization of Chemometric Modeling Approaches

To clarify the influence of key chemometric factors on the performance of fermentation process prediction models, four independent experimental subsections were designed in this research section, and the experimental variables were strictly controlled for each subsection to realize the independent analysis of a single factor. The first three subsections focused on investigating the individual effects of wavelength selection, pretreatment method selection and outlier elimination selection on modeling performance, respectively. A single-variable control strategy was adopted for these three subsections: in the wavelength selection experiment, only the wavelength range was adjusted for model construction without dividing the dataset into a calibration set and a validation set; the same experimental protocol was applied to the pretreatment method selection and outlier elimination selection experiments, where only the target factor (pretreatment method or outlier elimination strategy) was altered for modeling, with no division of calibration and validation sets. The sole objective of these three subsections was to independently identify the intrinsic influence of each single chemometric factor on model performance, and to screen out the optimal wavelength range, pretreatment method and outlier elimination strategy for fermentation process modeling. (In the outlier elimination subsection, the "appropriate secondary removal" in Figure 2 denotes a two-step procedure: initial spectral quality screening by the vendor's software, followed by secondary outlier removal based on prediction error threshold. The results section will provide a more detailed explanation).

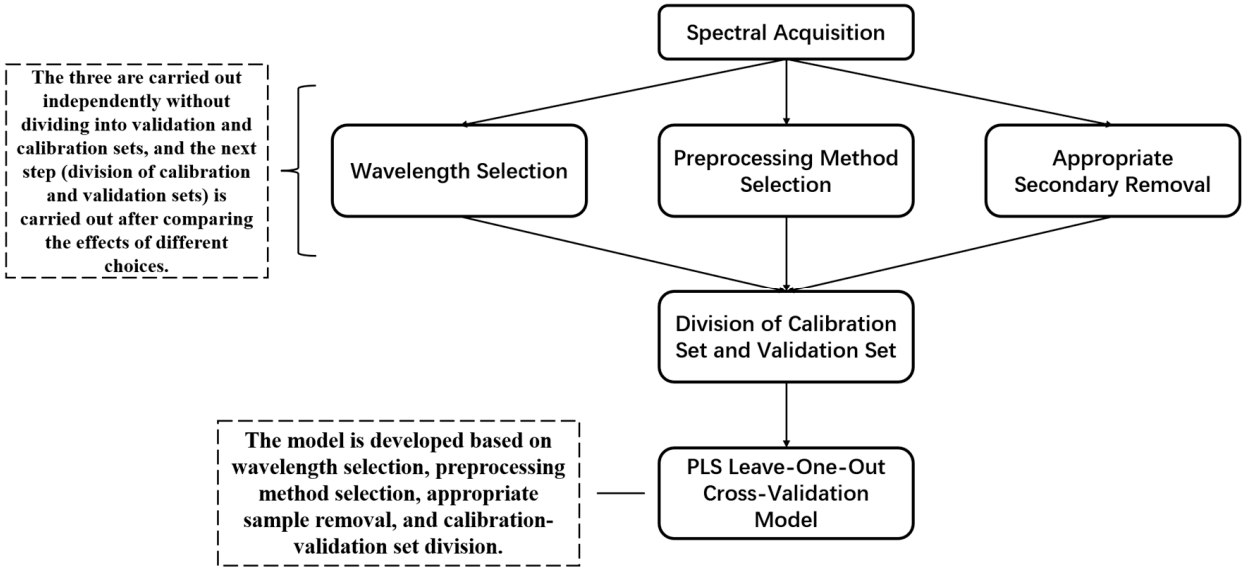


Figure 2. Flowchart of model construction and its design ideas.

The fourth subsection was designed to explore the influence of different division methods of calibration set and validation set on modeling performance under the condition of fixed sample size. Notably, this subsection was implemented on the basis of the optimal results obtained from the aforementioned three subsections: the modeling process adopted the screened optimal wavelength range, the most suitable pretreatment method and the optimal outlier elimination strategy. It is worth emphasizing that the first three subsections for factor screening were all evaluated by leave-one-out cross-validation (LOOCV) under a strict single-variable control strategy, rather than large-scale automatic optimization or massive combinatorial searching. Nested cross-validation was not employed in this stepwise screening process, since only a limited number of key factors were compared independently, thus leading to a low risk of overfitting and information leakage. This experimental design eliminated the interference of other chemometric factors, ensuring that the observed variation in model performance could be solely attributed to the difference in the division methods of calibration and validation sets, thus realizing the accurate evaluation of the effect of dataset division on the final modeling performance of the fermentation process prediction model. Figure 2 is a flow chart for modeling.

3. Results

3.1. Sensor Compatibility in Fermentation Scenarios

Different detector configurations have been employed across fermentation studies: Jin et al. [35] utilized transflectance, Kasemsumran et al. [36] employed transmission, and JM Andrés [37] applied diffuse reflectance. The present study simultaneously deployed both transflectance and transmission sensors in a parallel-loop configuration, enabling direct comparison of sensor performance in fermentation environments without mutual interference between flow cells. Offline glucose measurements were correlated with spectra acquired by each sensor type. Despite analyzing identical fermentation broth, the distinct detection principles generated characteristic spectral profiles (Figure 3): the 900–1300 nm region captured C–H functional group information, the 1400 nm band corresponded to strong water absorption, and O–H functional groups were represented in the 1300–1500 nm range.

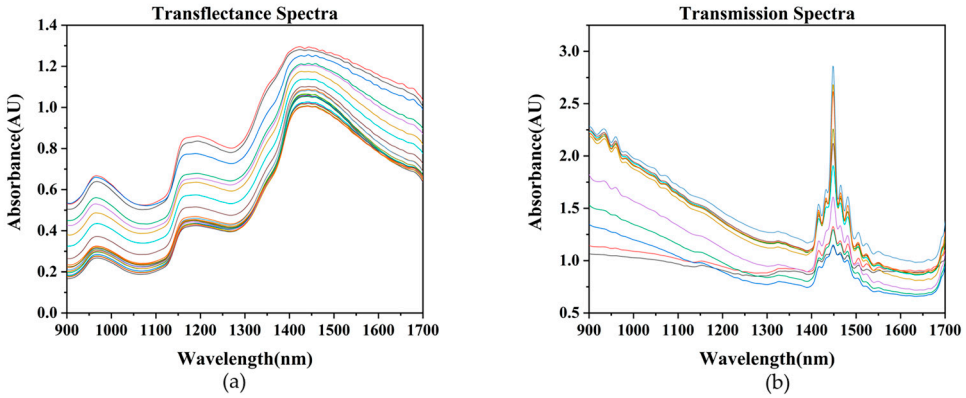


Figure 3. The collected original spectra. Transflectance Spectra (**a**), Transmission Spectra (**b**). Different colors are used to distinguish different samples, with no additional specific meaning.

As summarized in Table 2, the transflectance configuration demonstrated superior predictive performance compared to the transmission mode. Quantitatively, the transflectance model achieved a lower Root Mean Square Error of Calibration (RMSEC = 4.087 g/L) and Prediction (RMSEP = 8.512 g/L), coupled with a higher determination coefficient for validation (R^2_{val} = 0.988). Most notably, the RPD value for the transflectance model (3.735) significantly outperformed the transmission model (2.369), indicating a high level of analytical reliability. These results, supported by the tighter clustering of data points observed in the predicted-versus-reference plots (Figure 4), confirm that the transflectance setup offers enhanced fitting precision and more robust error control for glucose monitoring. Liu [38] previously reported that reflectance mode yielded higher detection accuracy than transmission when comparing spectroscopic approaches. The transflectance mode employed here—combining transmission and reflection principles—offers enhanced sensitivity, making it particularly advantageous for complex fermentation matrices where multiple analytes coexist.

Table 2. Performance of the PLS model.

Glucose Model	RMSEC	RMSEP	RMSECV	R^2_{cal}	R^2_{val}	R^2_{cv}	RPD	SD	RER	LVs
Transflectance	4.087	8.512	10.062	0.974	0.988	0.874	3.735	31.79	9.28	2
Transmission	5.972	9.443	7.021	0.954	0.981	0.937	2.369	22.37	4.87	1

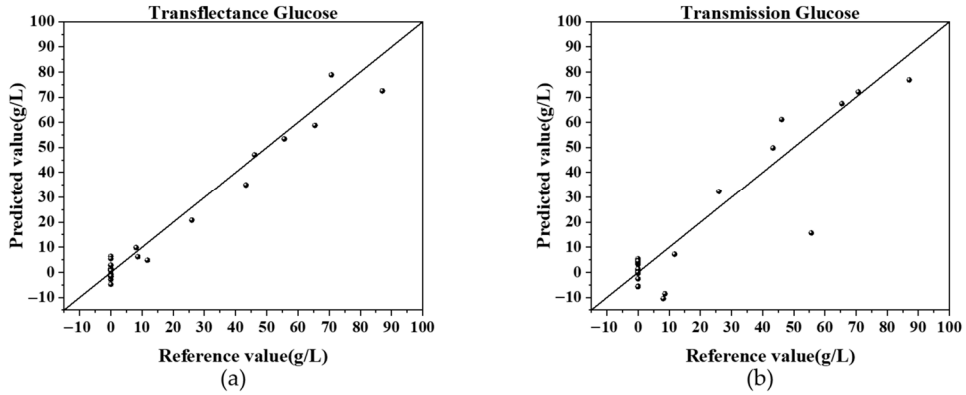


Figure 4. Predicted vs. reference glucose concentration plots for PLS regression models. The diagonal line represents the ideal prediction (predicted = reference). (**a**) Transflectance acquisition method; (**b**) Transmission acquisition method. X-axis: reference glucose concentration (g/L); Y-axis: predicted glucose concentration (g/L).

Glucose samples at 0 g/L in this study are an objective reflection of the actual fermentation process, rather than anomalous data caused by sensor baseline deviations or repeated measurements. These samples were collected in the late fermentation stage, where glucose (the carbon source) was completely consumed by microorganisms with fermentation still in progress, thus truly representing the actual state of the fermentation system at this phase. To comprehensively compare the adaptability of the two sensors in the fermentation environment, samples covering the entire fermentation cycle were collected; the inclusion of samples from the glucose-depleted stage is a necessary experimental design for evaluating their full-process detection performance.

This data distribution reflects the comprehensive process monitoring capability of the system. While a concentration of 0 g/L is a challenging boundary condition for regression, its inclusion ensures that the model is calibrated across the entire physiological range of the fermentation cycle. Rather than introducing bias, these authentic in situ data points enhance the model's robustness and practical utility, allowing for an objective evaluation of the sensors' performance during the critical substrate-depletion phase. By incorporating these samples (representing less than 5% of the total dataset), the resulting calibration more accurately mirrors the real-world operational environment, ensuring the model remains reliable even as the carbon source is exhausted.

3.2. Impact of Medium Composition on Spectral Acquisition

Fang et al. [39] modeled the NIR spectra of hemoglobin in biological solutions under different water backgrounds and found that changes in water content significantly affected the robustness of the NIR model, which indirectly shows that changes in liquid composition affect near-infrared acquisition. Peptone and yeast extract, common medium constituents, modify both coloration and turbidity. Given that both transmission and transflectance sensors acquire liquid-phase spectra, medium composition and turbidity inevitably affect spectral characteristics. To test this hypothesis, fermentations were conducted using two media formulations—synthetic medium and complex medium—with glucose models constructed for each to assess spectral acquisition under differing matrix conditions. The complex medium contained supplemental peptone and yeast extract relative to the synthetic formulation, introducing compositional complexity and elevated turbidity. Model performance comparisons quantified the impact of these matrix variations on spectral data quality.

In synthetic culture media, the transmission mode demonstrated superior predictive performance compared to the transflectance mode. As shown in Table 3 for the model values, the transmission-based PLS model yielded lower root mean square errors for both calibration (RMSEC = 11.038 g/L) and prediction (RMSEP = 11.724 g/L), with coefficients of determination (R^2_{cal} and R^2_{val}) both reaching 0.979. The Range Error Ratio (RER) for the transmission mode was as high as 15.18, significantly exceeding that of the transflectance mode (8.35), which, combined with a Residual Prediction Deviation (RPD) of 6.118, underscores the model's excellent precision and practical utility in this environment. Conversely, in complex medium environments, the transflectance sensor exhibited significantly higher robustness. As shown in Table 4, the transflectance model maintained near-unity determination coefficients (val = 0.993 and R^2_{val} = 0.971) and a relatively low RMSEC of 2.162 g/L. In a more complex medium environment, the transflectance sensor exhibits better performance. For example, its RPD value is 6.722 and RER is 12.81. In stark contrast, the performance of the transmission mode collapsed under these conditions, characterized by a poor R^2_{cal} of only 0.594, an RPD of 1.81, and an RER of 3.458, indicating its failure to provide accurate or stable predictions when faced with the high turbidity and chemical complexity of the complex medium. This is also consistent with the results presented

in Model Dispersion Figure 5. Hong et al. [40] noted that transmission produces weak signals for opaque or highly scattering samples. In this experiment, complexifying the matrix by incorporating peptone and yeast extract revealed that transmission experienced stronger interference, not only validating that transmission is more suitable for simple, clear samples but also demonstrating that complex culture media exert negative effects on transmission-acquired spectra.

Table 3. Performance of PLS models established in synthetic culture media environments.

Synthetic Medium (Glucose)	RMSEC	RMSEP	RMSECV	R^2_{cal}	R^2_{val}	R^2_{cv}	RPD	SD	RER	LVs
Transflectance	19.282	21.315	26.382	0.937	0.959	0.883	3.365	71.24	8.35	6
Transmission	11.038	11.724	14.624	0.979	0.979	0.963	6.118	71.72	15.18	5

Table 4. Performance of PLS models established in complex medium environments.

Complex Medium (Glucose)	RMSEC	RMSEP	RMSECV	R^2_{cal}	R^2_{val}	R^2_{cv}	RPD	SD	RER	LVs
Transflectance	2.162	3.198	9.524	0.993	0.971	0.884	6.722	21.49	12.81	11
Transmission	25.934	11.856	29.581	0.134	0.594	0.002	1.81	21.45	3.458	2

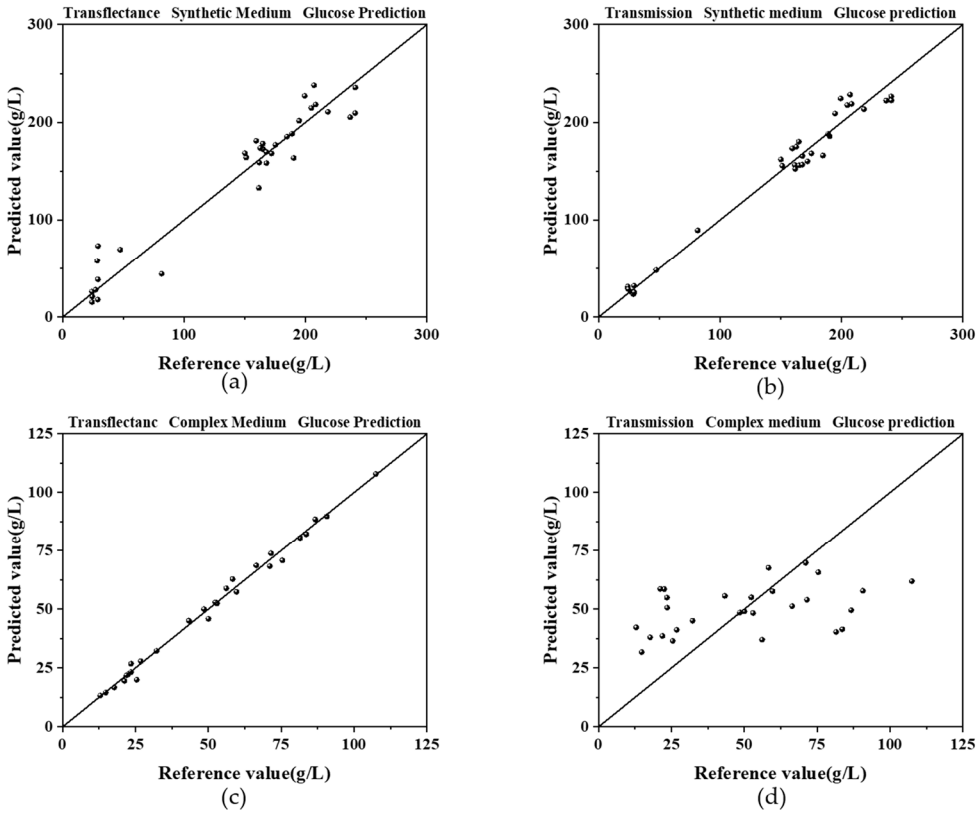


Figure 5. Predicted vs. reference glucose concentration plots for PLS regression models under two culture medium conditions. The diagonal line represents the ideal prediction (predicted = reference). (**a**) Transflectance model in synthetic medium; (**b**) Transmission model in synthetic medium; (**c**) Transflectance model in complex medium; (**d**) Transmission model in complex medium. X-axis: reference glucose concentration (g/L); Y-axis: predicted glucose concentration (g/L).

3.3. Model Optimization

3.3.1. Wavelength Selection

The spectral bands acquired by NIR contain highly complex information, and no fixed optimal wavelength band range exists for detection parameters. In this experiment, 48 samples were used for modeling, and multiple wavelength band ranges were manually selected to study the impact of different wavelength band selections on model performance. The five bands selected as shown in Table 5, Wavelength 5 (1550–1700 nm) proved optimal, exhibiting the lowest RMSEC (8.265 g/L), R^2_{cal} reaching 0.909, as well as excellent cross-validation root mean square error RMSECV value (10.536 g/L) and R^2_{cv} (0.852), demonstrating outstanding model accuracy and predictive capability. Wavelength 2 (900–1300 nm) and Wavelength 4 (1400–1550 nm) performed poorly. For example, Wavelength 4 has a calibration coefficient of determination (R^2_{cal}) of 0.839 and a cross-validation coefficient of determination (R^2_{cv}) of 0.761, indicating insufficient information content in this range. Scatter plot (Figure 6) validation confirmed that Wavelength 5 prediction points aligned most closely with the diagonal line, showing high consistency, whereas Wavelength 2 exhibited dispersed scatter points with large errors. These results demonstrate that the narrow band 1550–1700 nm reduces redundancy, enhances characteristic features, and significantly improves model performance.

Table 5. Performance of PLS models with different wavelength selections.

	RMSEC	RMSECV	R^2_{cal}	R^2_{cv}	LVs
Wavelength 1 (900–1700 nm)	8.982	10.415	0.893	0.856	5
Wavelength 2 (900–1300 nm)	8.706	13.76	0.899	0.754	8
Wavelength 3 (1400–1700 nm)	9.533	10.852	0.879	0.844	4
Wavelength 4 (1400–1550 nm)	10.997	13.525	0.839	0.761	5
Wavelength 5 (1550–1700 nm)	8.265	10.536	0.909	0.852	6

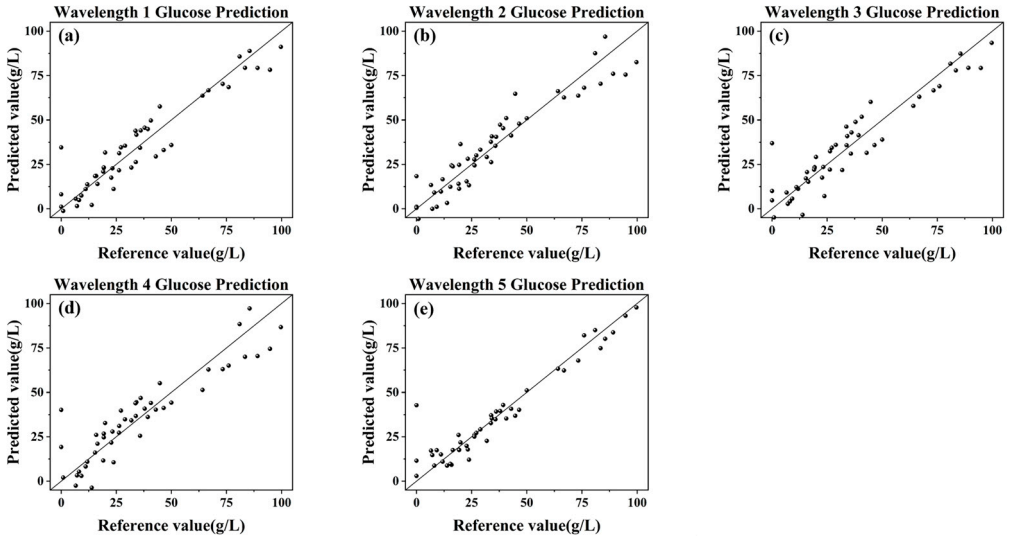


Figure 6. Predicted vs. reference glucose concentration plots for PLS regression models under different wavelength selections. The diagonal line represents the ideal prediction (predicted = reference). (**a**) Wavelength 1; (**b**) Wavelength 2; (**c**) Wavelength 3; (**d**) Wavelength 4; (**e**) Wavelength 5. X-axis: reference glucose concentration (g/L); Y-axis: predicted glucose concentration (g/L).

Full-range prediction plots (Figure 7) similarly reflected differences among wavelength selections: Wavelength 2 (900–1300 nm), not being a characteristic band, showed excessive discrepancies between predicted and actual values; Wavelength 5 (1550–1700 nm), despite achieving the best model parameters, exhibited predicted values closest to actual values

but demonstrated reduced stability with fluctuations throughout the prediction range; wavelength 1 (900–1700 nanometers) has a balanced prediction accuracy with stability over the entire prediction range, but there is a deviation between its predicted values and the actual values. Di Egidio et al. [41] observed several absorption bands in wine fermentation, such as 7139 cm^{-1} (1400 nm) and 5955–5866 cm^{-1} (around 1700 nm), corresponding to O–H and C–H bonds. Since most chemical bonds constituting glucose fall within these regions, the model established using Wavelength 2 (900–1300 nm) performed poorest due to deviation from characteristic absorption regions. Although Wavelength 5 (1550–1700 nm) established a superior model, it exhibited insufficient stability in full-range prediction plots. Therefore, wavelength selection for modeling must encompass characteristic absorption bands while avoiding excessively narrow ranges.

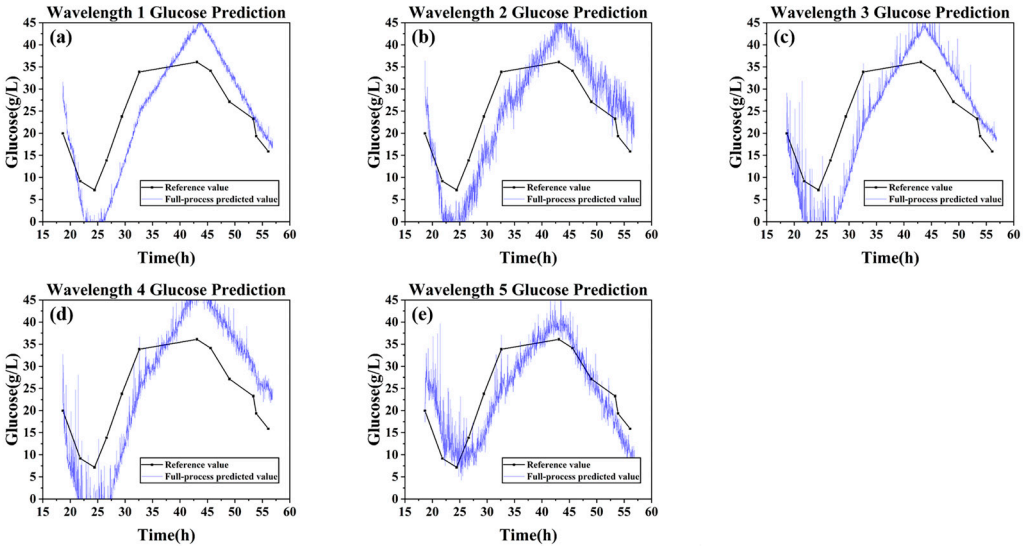


Figure 7. Time-series prediction of glucose concentration during fermentation using PLS regression models under different wavelength selections. (**a**) Wavelength 1; (**b**) Wavelength 2; (**c**) Wavelength 3; (**d**) Wavelength 4; (**e**) Wavelength 5. The x-axis represents fermentation time (h), and the y-axis represents glucose concentration (g/L). The black solid line denotes the reference glucose value, while the blue solid line denotes the full-process predicted glucose value.

3.3.2. Preprocessing Method Selection

The high sensitivity of NIR to O-H bonds means that the spectral information is dominated by water absorption, which often overlaps with the weaker C-H and O-H vibrations of glucose. Additionally, matrix effects in liquid samples introduce non-linear variations and baseline drifts.

This experiment aims to investigate the impact of selecting different preprocessing methods on modeling, using 48 samples. Calibration set parameters (RMSEC, R^2_{cal}) in the table reveal (Table 6) that most basic preprocessing methods (such as moving average smoothing, Savitzky–Golay (SG) smoothing, and standardization) yielded RMSEC values concentrated in the 8.9–9 g/L range with R^2_{cal} approximately 0.892, indicating moderate fitting performance. However, normalization, SNV, and SG first-order differentiation improved the accuracy of the model. For example, the RMSEC after normalization preprocessing decreased to 2 g/L, but the model after SG second-order differentiation reduced the accuracy, with its RMSEC increasing to 10.2 g/L.

Table 6. Model performance of PLS models established after different preprocessing methods.

	RMSEC	RMSECV	R^2_{cal}	R^2_{cv}	LVs
Moving average smoothing	9.002	10.405	0.892	0.857	5
SG smoothing	8.986	10.416	0.892	0.856	5
Standardization	8.982	10.415	0.893	0.857	5
Normalization	2.022	12.978	0.944	0.816	10
MSC	8.292	12.589	0.908	0.792	6
SNV	2.022	12.978	0.994	0.816	10
SG first-order differential	6.663	10.012	0.941	0.867	5
SG second-order differential	10.205	16.171	0.861	0.66	4

Cross-validation metrics (RMSECV and R^2_{cv}) reflect the model's predictive capability for new samples. Basic preprocessing methods (e.g., moving average smoothing and SG smoothing) maintained RMSECV values of 10.4 g/L and R^2_{cv} around 0.85, demonstrating balanced generalization performance. In contrast, normalization, multiplicative scatter correction (MSC), and standard normal variate (SNV) methods exhibited reduced R^2_{cv} (0.79–0.8) due to excessive noise suppression or information loss, compromising generalization ability. Although the SG derivative exhibits good calibration performance, its generalization ability decreases significantly: SG first-order derivative yielded RMSECV of 10.012 g/L and R^2_{cv} of 0.867, while SG second-order derivative performed even worse with RMSECV reaching 16.171 g/L and R^2_{cv} plummeting to 0.66, indicating severe overfitting to the calibration set and substantial prediction errors for new samples. This can also be seen in the scatter plot (Figure 8).

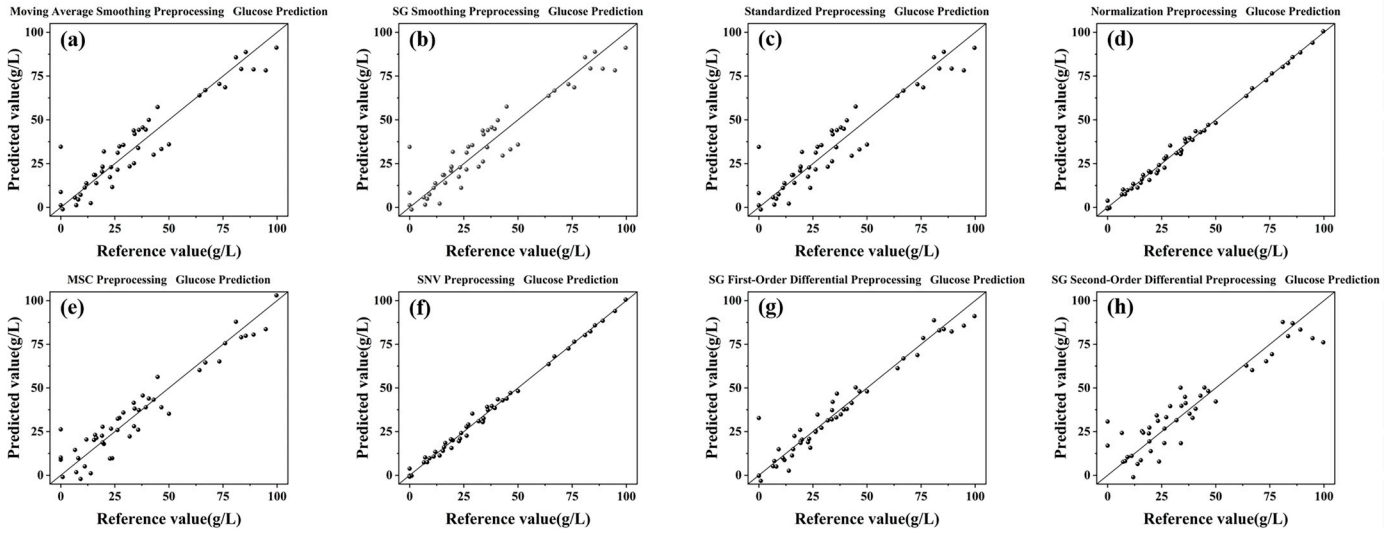


Figure 8. Predicted vs. reference glucose concentration plots for PLS regression models established with different spectral preprocessing methods. The diagonal line represents the ideal prediction (predicted = reference). (**a**) Moving average smoothing; (**b**) Savitzky–Golay (SG) smoothing; (**c**) Standardization; (**d**) Normalization; (**e**) Multiplicative scatter correction (MSC); (**f**) Standard normal variate (SNV); (**g**) First-order SG differentiation; (**h**) Second-order SG differentiation. X-axis: reference glucose concentration (g/L); Y-axis: predicted glucose concentration (g/L).

Prediction plots (Figure 9) visually demonstrate the alignment between predicted curves (blue lines) and reference values (black lines). Basic preprocessing methods such as moving average smoothing and standardization produced prediction curves that followed the overall trend of reference values with acceptable fluctuations. Conversely, the Savitzky–Golay (SG) second-order derivative prediction curve exhibited severe oscillations

and substantial deviations from reference values, confirming its poor generalization capability. Preprocessing method selection requires balancing calibration and generalization performance through comprehensive evaluation of prediction plots and numerical metrics. Based on this analysis, it is determined that basic methods such as the moving average smoothing method are most suitable for the development of glucose models.

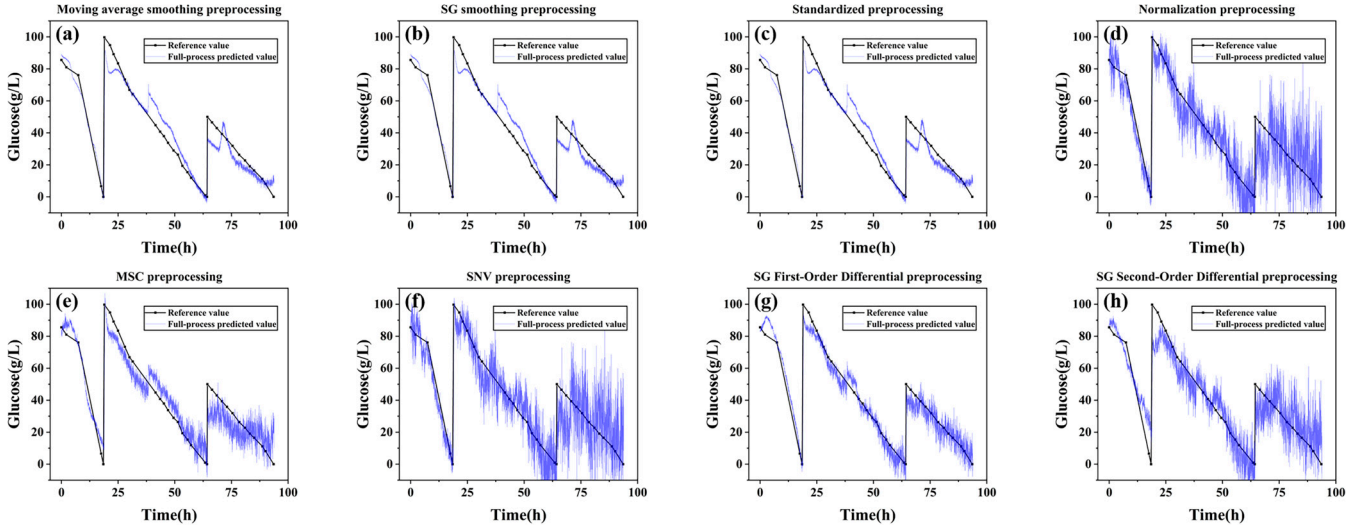


Figure 9. Time-series prediction of glucose concentration during fermentation using PLS regression models with different spectral preprocessing methods. (**a**) Moving average smoothing; (**b**) Savitzky–Golay (SG) smoothing; (**c**) Standardization; (**d**) Normalization; (**e**) Multiplicative scatter correction (MSC); (**f**) Standard normal variate (SNV); (**g**) First-order SG differentiation; (**h**) Second-order SG differentiation. The x-axis represents fermentation time (h), and the y-axis represents glucose concentration (g/L). The black solid line denotes the reference glucose value, while the blue solid line denotes the full-process predicted glucose value.

In liquid NIR analysis, the signal-to-noise ratio is significantly challenged by the strong absorption of water and complex matrix effects. While advanced techniques such as Orthogonal Signal Correction (OSC) or Orthogonal Projection can potentially isolate glucose-specific signals by removing Y-unrelated variances, they often require higher computational resources and specific algorithm implementation. In this study, the observed performance of different methods reflects these interferences: the failure of SG derivatives underscores the sensitivity of high-order transformations to water-related background noise. Consequently, Moving Average smoothing was identified as the most robust and practical choice for this application, balancing noise suppression with spectral integrity.

3.3.3. Outlier Spectral Sample Removal

Online near-infrared spectral data are first objectively screened by the vendor's software using built-in spectral diagnostic standards (baseline, signal-to-noise ratio, etc.). On the basis of the manufacturer-certified spectrally valid data, a secondary outlier screening via prediction error threshold was performed as a retrospective sensitivity analysis. The research objective was to investigate the impact of measurement noise and "high-residual" samples on the stability and theoretical performance limits of the PLS model.

Based on this objective, the experiment utilized 39 samples. After establishing the initial PLS model, samples were sequentially removed based on error thresholds (Removal 1 through 5 corresponding to errors > 20, 9.5, 8, 7, and 5, eliminating 3, 7, 12, 14, and 19 samples, respectively). It should be noted that since this outlier identification was conducted on the full dataset rather than strictly within calibration folds, the resulting

performance metrics may exhibit an optimistic bias due to circular validation. However, this analysis remains valuable for understanding model behavior under different data-quality constraints.

Statistical indicators revealed that RMSEC initially decreased sharply then increased (Table 7): from 9.218 g/L at No removal to a minimum of 1.441 g/L at Removal 3, but rising to 2.398 g/L at Removal 5 due to excessive sample reduction. RMSECV also follows a pattern of first decreasing and then increasing, reaching its minimum value of 4.874 g/L when 3 is removed. R^2_{cal} increased steadily with sample removal, reaching 0.996 at removal 3 and removal 4, but slightly decreased to 0.989 at removal 5, indicating an overall improvement in calibration set fitting performance. R^2_{cv} gradually increased, peaking at 0.962 at removal 3, demonstrating that removing high-error samples significantly enhanced the model's predictive capability for unknown samples. The scatter plot of predicted values versus reference values (Figure 10) also demonstrates this result.

Table 7. Model performance of PLS models established after removing different numbers of samples.

	RMSEC	RMSECV	R^2_{cal}	R^2_{cv}	LVs
No removal	9.218	13.473	0.867	0.725	6
Removal 1 (error > 20, 3 samples removed)	4.488	6.625	0.966	0.928	7
Removal 2 (error > 9.5, 7 samples removed)	2.442	6.343	0.991	0.937	10
Removal 3 (error > 8, 12 samples removed)	1.441	4.874	0.996	0.962	10
Removal 4 (error > 7, 14 samples removed)	1.499	6.225	0.996	0.937	10
Removal 5 (error > 5, 19 samples removed)	2.398	6.452	0.989	0.928	7

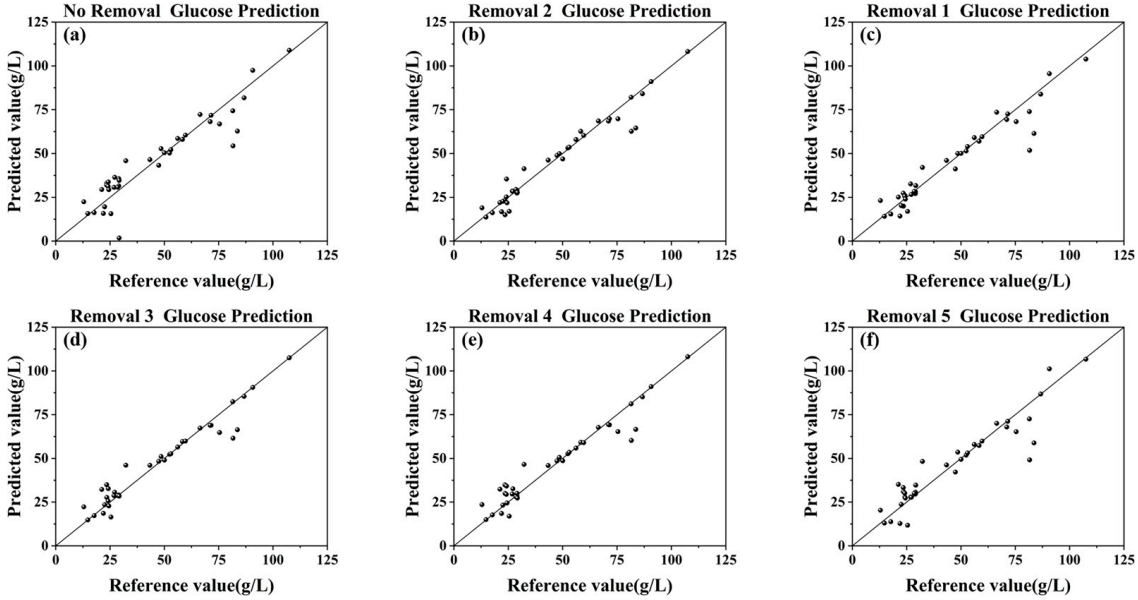


Figure 10. Predicted vs. reference glucose concentration plots for PLS regression models established after removing different numbers of samples. The diagonal line represents the ideal prediction (predicted = reference). (**a**) No removal; (**b**) Removal 1; (**c**) Removal 2; (**d**) Removal 3; (**e**) Removal 4; (**f**) Removal 5. X-axis: reference glucose concentration (g/L); Y-axis: predicted glucose concentration (g/L).

The six sets of glucose prediction plots (Figure 11) showed progressive improvement in alignment between predicted curves and reference values with increasing sample removal, though Removal 5 exhibited performance decline due to elimination of nearly half the sample population. Moderate removal of high-error outlier samples (e.g., Removal 3, eliminating 12 samples) effectively optimized model performance by balancing calibration accuracy and generalization capability. However, excessive removal (e.g., Removal 5) may elevate calibration error due to insufficient sample size. As a retrospective sensitivity analysis, this defines the model's theoretical performance limits. Although this filtering introduces an optimistic bias by utilizing known reference values, it provides a valuable diagnostic benchmark for calibration maintenance. In practice, thresholds must be carefully selected to balance fitting accuracy with sufficient sample size to ensure model generalization.

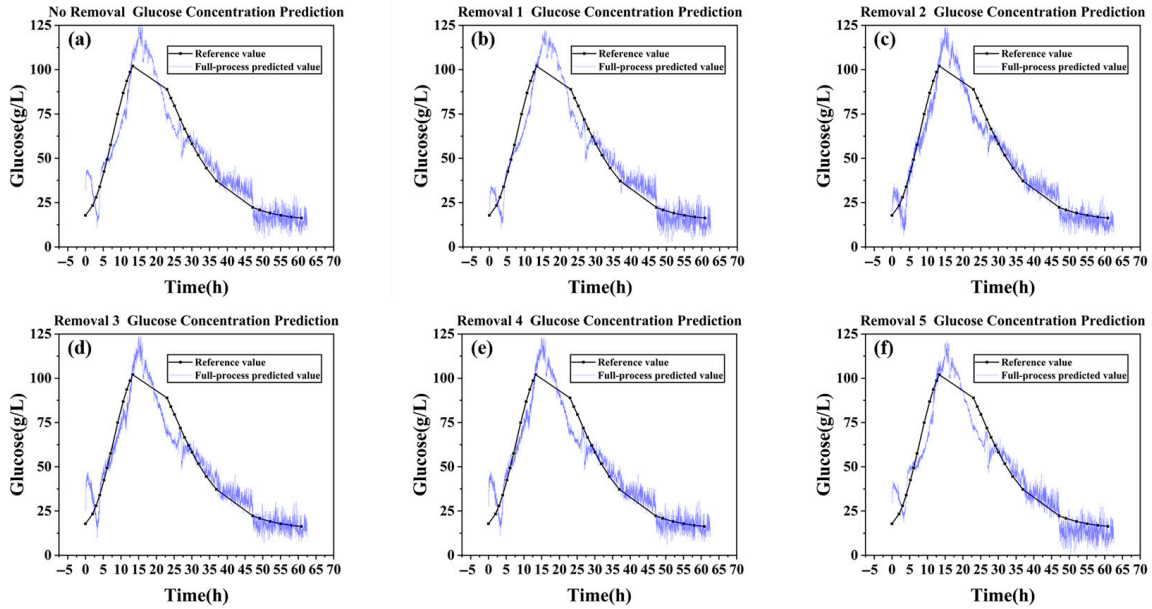


Figure 11. Time-series prediction of glucose concentration during fermentation using PLS regression models after removing different numbers of samples. (**a**) No removal; (**b**) Removal 1; (**c**) Removal 2; (**d**) Removal 3; (**e**) Removal 4; (**f**) Removal 5. The x-axis represents fermentation time (h), and the y-axis represents glucose concentration (g/L). The black solid line denotes the reference glucose value, while the blue solid line denotes the full-process predicted glucose value.

3.3.4. Calibration and Validation Set Partitioning

After selecting the wavelength, preprocessing method, and removing samples, experiments were conducted using 46 samples. A strategy of randomly dividing the calibration set/validation set was employed, which is specifically designed to avoid the interference of temporal autocorrelation and other irrelevant variables. This enables a dedicated univariate study on the impact of the calibration set/validation set ratio division on the performance of near-infrared (NIR) modeling. When the calibration set is sufficiently large (e.g., Validation 1: 36 calibration samples, 10 validation samples), both RMSEC (4.121 g/L) and root mean square error of validation (RMSEP; 3.476 g/L) remain low, R^2_{cal} and coefficient of determination for validation (R^2_{val}) both approach unity, and ratio of performance to deviation (RPD; 8.038 > 5) indicates robust predictive capability (Table 8; Figure 12). Under these conditions, the calibration set effectively captures spectral-concentration relationships, while the validation set reliably assesses generalization performance. Correspondingly, predicted values in Validation 1's (Figure 13a) graphical representation exhibit localized fluctuations but maintain strong overall agreement with reference values. As the number

of calibration samples gradually decreases (for example, in Validation 3: 27 calibration samples and 19 validation samples), the number of samples in the calibration set and the validation set becomes quite similar, and the RPD value also drops to 3.48. However, it still has a certain predictive ability, but the robustness of the prediction has decreased, which can be reflected in part C of Figure 13. To try an extreme case, a fifth division was performed (12 calibration samples and 34 validation samples). The excessively small calibration set caused R^2_{cv} to drop to 0.887. Although the predicted trend is consistent, there is still a certain deviation from the reference value, and the prediction fluctuation has also increased. This indicates that the model performance gradually decreases from having high-precision quantitative prediction ability. The corresponding charts show that the predicted values are unstable, indicating that insufficient calibration samples will hinder the full learning of the spectrum-glucose correlation, leading to the failure of calibration and validation.

Table 8. Evaluate the performance of PLS models established at different ratios of training and validation sets.

	RMSEC	RMSEP	RMSECV	R^2_{cal}	R^2_{val}	R^2_{cv}	RPD	SD	RER	LVs
Validation 1 (36 calibration, 10 validation)	4.121	3.476	7.644	0.976	0.992	0.919	8.038	27.94	27.26	7
Validation 2 (32 calibration, 14 validation)	2.617	6.618	5.554	0.99	0.949	0.956	4.275	28.29	14.32	9
Validation 3 (27 calibration, 19 validation)	2.248	7.483	5.196	0.993	0.937	0.965	3.486	26.08	12.66	9
Validation 4 (20 calibration, 26 validation)	2.111	6.96	6.845	0.994	0.953	0.933	3.963	27.58	13.61	9
Validation 5 (12 calibration, 34 validation)	0.506	7.288	10.638	0.999	0.935	0.887	3.521	25.66	13	9

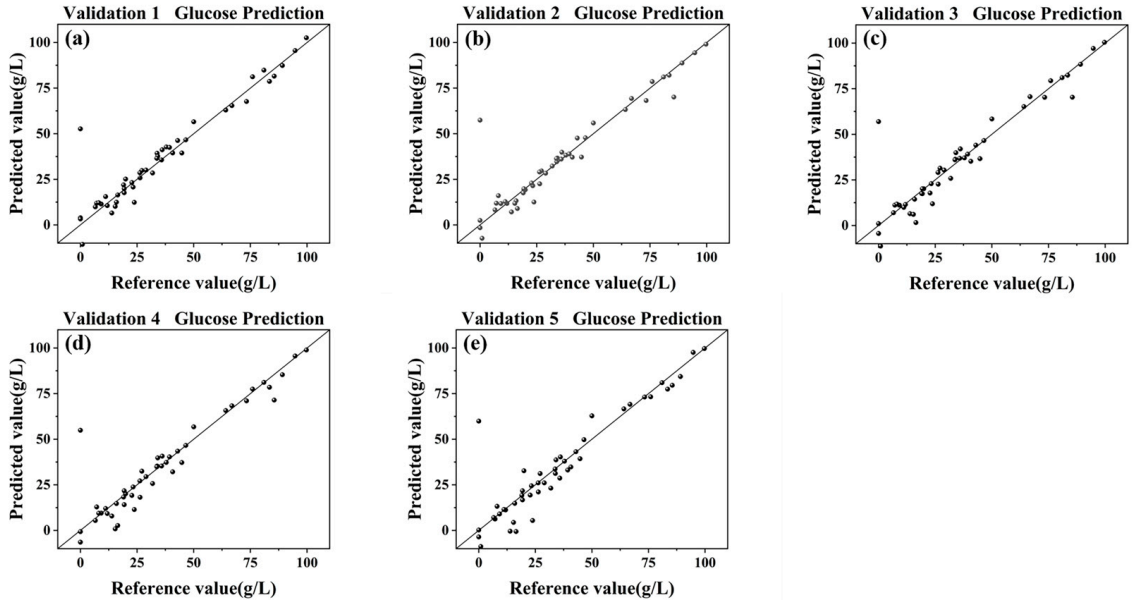


Figure 12. Predicted vs. reference glucose concentration plots for PLS regression models constructed using different calibration and validation set partitioning methods. The diagonal line represents the ideal prediction (predicted = reference). (**a**) Validation 1; (**b**) Validation 2; (**c**) Validation 3; (**d**) Validation 4; (**e**) Validation 5. X-axis: reference glucose concentration (g/L); Y-axis: predicted glucose concentration (g/L).

The partitioning of calibration and validation sets must balance learning and assessment functions. With a fixed total sample size, the calibration set requires sufficient

magnitude (e.g., 60–70% allocation) to stabilize model parameters and avoid underfitting, while the validation set necessitates moderate size (e.g., 30–40%) to authentically reflect generalization capacity and prevent overfitting. Excessive reduction in calibration samples or disproportionate expansion of validation samples triggers error amplification and predictive collapse. In this study, Validation 1's partitioning (36 calibration, 10 validation samples) achieves optimal performance by harmonizing both objectives, providing empirical evidence for rational sample allocation in NIR modeling. This conclusion can serve as a fundamental sample design principle for NIR modeling of fermentation processes that incorporate time-series characteristics and cross-batch generalization.

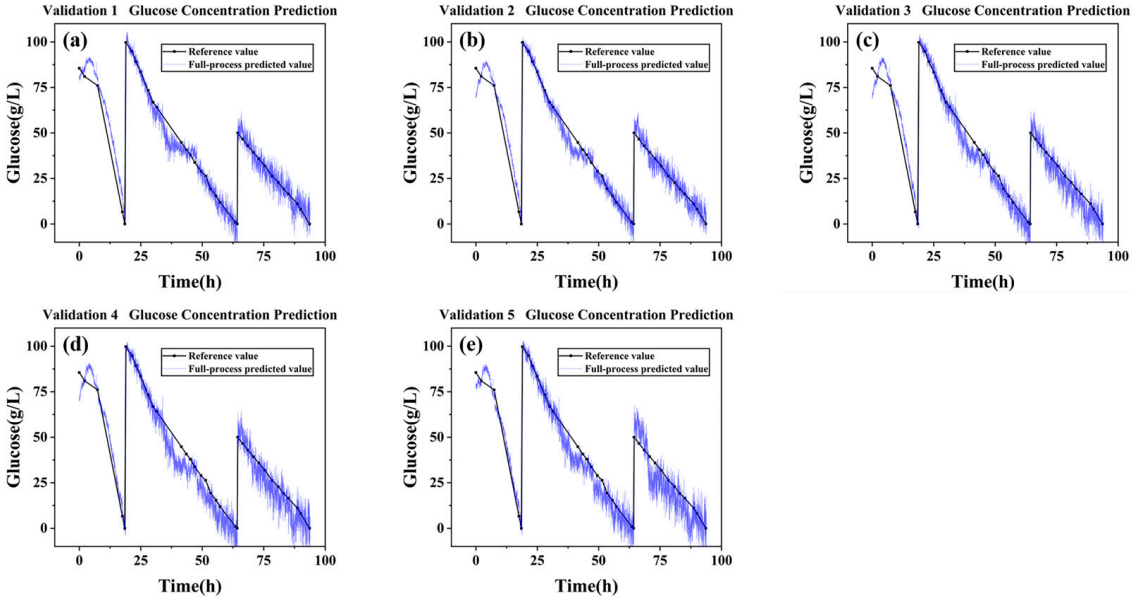


Figure 13. Time-series prediction of glucose concentration during fermentation using PLS regression models trained with five different validation set splitting methods. (**a**) Validation 1; (**b**) Validation 2; (**c**) Validation 3; (**d**) Validation 4; (**e**) Validation 5. The x-axis represents fermentation time (h), and the y-axis represents glucose concentration (g/L). The black solid line denotes the reference glucose value, while the blue solid line denotes the full-process predicted glucose value.

4. Discussion

The performance divergence observed between the sensors in this study highlights the critical role of light–matter interaction in heterogeneous biological systems. In clear synthetic media, both sensors perform effectively as photon propagation follows the linear Beer–Lambert law. However, the transition to complex media introduces significant Mie scattering caused by suspended organic particles and macromolecules. This scattering leads to severe signal attenuation in transmission mode, as photons are diverted away from the detector. In contrast, the transflectance sensor maintains robust performance by utilizing an integrated reflector to "recycle" backscattered photons, ensuring a consistent interaction volume near the probe tip. This aligns with the comparative analysis by Jeon et al. [42], who demonstrated that while transmission is ideal for clear solutions, reflected/transflected paths are superior for compensating scattering effects in highly turbid biological matrices.

The selection of the 1550–1700 nm spectral region as the "golden window" for glucose monitoring is based on a strategic balance between molecular specificity and solvent interference. In aqueous fermentation broths, the massive O–H absorption band centered at 1450 nm creates an optical "blind spot" that masks the weaker signals of solutes. The failure of the 1400–1550 nm region in this study confirms the deleterious effect of this water peak. In contrast, the 1550–1700 nm wavelength band can distinguish the C–H bond vibrations of

glucose with higher specificity, as discovered by M Yang et al. [43], where the peaks around 1605 nm and 1706 nm are more suitable for glucose detection. While the combination band at 2100–2300 nm often provides stronger absorptivity, its practical application in industrial fermentation is limited by extremely high water absorbance and shallow penetration [44], making the 1550–1700 nm first overtone region a more reliable choice for online monitoring in large-scale reactors [3]. Regarding data processing, the results underscore the danger of mathematical over-optimization in chemometrics, particularly when working with a relatively small dataset. For example, high-order preprocessing such as the Savitzky–Golay second derivative will fail in cross-validation. As cautioned by Rinnan et al. [45] in their comprehensive review of NIR preprocessing, numerical differentiation inherently amplifies high-frequency noise, which in dynamic fermentation environments (e.g., from agitation or micro-bubbles) is often mistakenly captured as chemical information. These findings serve as a cautionary note that achieving high calibration accuracy through complex preprocessing should not be prioritized over model generalizability in small-scale studies. Moving Average Smoothing proved more appropriate by suppressing noise without distorting the underlying peak shapes. Similarly, outlier management must be governed by physical justifications—such as identifying spectral artifacts caused by transient bubbles—rather than solely chasing minimized error metrics. An aggressive removal of data may lower numerical errors but results in a "fragile" model that lacks the statistical representativeness necessary for real-world process variability.

The experimental platform developed in this study constitutes a prototype with preliminarily demonstrated feasibility for Process Analytical Technology (PAT) in laboratory-scale biomanufacturing. By combining the structural benefits of transflectance probes with specific band selection and conservative data processing, this work demonstrates the feasibility of reliable online glucose monitoring in complex, turbid fermentation media at the laboratory scale. The prediction accuracy obtained in this study can meet the preliminary requirements for coarse-to-medium control in laboratory fermentation scenarios. This method exhibits promising scalability and transfer potential to other typical aerobic fermentation processes (e.g., organic acid or amino acid production). Overall, this study explores a feasible technical route for the real-time monitoring of laboratory fermentation processes, providing a reference for the subsequent digital transformation and automated intelligent monitoring of fermentation systems.

This study has certain limitations that should be acknowledged regarding the generalizability of its findings. Some modeled samples are individual fermentation batches, so they do not consider the inherent batch-to-batch variations common in industrial biomanufacturing, which may affect the stable performance of this method in various production batches. In addition, the current work focuses on a single microbial strain and glucose as the sole target analyte, with no validation for other fermentation substrates, industrial strains or key analytes (e.g., metabolic byproducts, biomass), limiting the immediate applicability of the developed method to more diverse fermentation systems.

Subsequent research will address these limitations to further verify the method's robustness and broad applicability. Multi-batch fermentation experiments will be conducted to incorporate batch-to-batch variability and improve the model's industrial relevance. The transflectance NIR platform will also be tested on alternative substrates, different industrially relevant microbial strains, and additional key fermentation analytes, to expand its scope for practical biomanufacturing applications.

Author Contributions: Conceptualization, G.S. and Y.L.; Methodology, S.Y.; Validation, S.Y. and Z.L.; Formal Analysis, S.Y. and Z.L.; Investigation, S.Y. and J.T.; Data Curation, S.Y. and Z.L.; Writing, S.Y. and Z.L.; Supervision, F.X., Y.L. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Wuxi Industrial Innovation Research Institute Pilot Technology Pre-research Project (XD24024), Jiangsu Funding Program for Excellent Postdoctoral Talent (2024ZB371), Jiangsu Basic Research Center for Synthetic Biology (Grant No. BK20233003), and the National Natural Foundation of China (32172174).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Zhikai Liu was employed by the company Wuxi Institute for Specialized Nutrition and Health Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Sharma, R.; Garg, P.; Kumar, P.; Bhatia, S.K.; Kulshrestha, S. Microbial fermentation and its role in quality improvement of fermented foods. *Fermentation* **2020**, *6*, 106. [CrossRef]
2. Siddiqui, S.A.; Erol, Z.; Rugji, J.; Taşçı, F.; Kahraman, H.A.; Toppi, V.; Musa, L.; Di Giacinto, G.; Bahmid, N.A.; Mehdizadeh, M. An overview of fermentation in the food industry-looking back from a new perspective. *Bioresour. Bioprocess.* **2023**, *10*, 85. [CrossRef] [PubMed]
3. Yang, C.; Lingli, C.; Meijin, G.; Xu, L.; Jinsong, L.; Xiaofeng, L.; Zhongbing, C.; Xiaojun, T.; Haoyue, Z.; Xiwei, T. Application of near-infrared spectroscopy technology in the complex fermentation system to achieve high-efficiency production. *Bioresour. Bioprocess.* **2021**, *8*, 96. [CrossRef] [PubMed]
4. Vann, L.; Layfield, J.B.; Sheppard, J.D. The application of near-infrared spectroscopy in beer fermentation for online monitoring of critical process parameters and their integration into a novel feedforward control strategy. *J. Inst. Brew.* **2017**, *123*, 347–360. [CrossRef]
5. Zhu, X.; Rehman, K.U.; Wang, B.; Shahzad, M. Modern soft-sensing modeling methods for fermentation processes. *Sensors* **2020**, *20*, 1771. [CrossRef]
6. Chai, W.Y.; Teo, K.T.K.; Tan, M.K.; Tham, H.J. Fermentation process control and optimization. *Chem. Eng. Technol.* **2022**, *45*, 1731–1747. [CrossRef]
7. Xia, J.; Long, D.; Chen, M.; Chen, A. Optimization of fermentation processes in intelligent biomanufacturing: On online monitoring, artificial intelligence, and digital twin technologies. *Sheng Wu Gong Cheng Xue Bao Chin. J. Biotechnol.* **2025**, *41*, 1179–1196. [CrossRef]
8. Manley, M. Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. *Chem. Soc. Rev.* **2014**, *43*, 8200–8214. [CrossRef]
9. Beć, K.B.; Grabska, J.; Huck, C.W. Miniaturized NIR spectroscopy in food analysis and quality control: Promises, challenges, and perspectives. *Foods* **2022**, *11*, 1465. [CrossRef]
10. Blanco, M.; Villarroya, I. NIR spectroscopy: A rapid-response analytical tool. *TrAC-Trends Anal. Chem.* **2002**, *21*, 240–250. [CrossRef]
11. Cozzolino, D. The ability of near infrared (NIR) spectroscopy to predict functional properties in foods: Challenges and opportunities. *Molecules* **2021**, *26*, 6981. [CrossRef] [PubMed]
12. Cervera, A.E.; Petersen, N.; Lantz, A.E.; Larsen, A.; Gernaey, K.V. Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnol. Prog.* **2009**, *25*, 1561–1581. [CrossRef]
13. Bec, K.B.; Grabska, J.; Huck, C.W. Near-infrared spectroscopy in bio-applications. *Molecules* **2020**, *25*, 2948. [CrossRef]
14. Ren, W.; Jiang, Q.; Qi, W. Research progress in near-infrared spectroscopy for detecting the quality of potato crops. *Chem. Biol. Technol. Agric.* **2025**, *12*, 32. [CrossRef]
15. Wang, H.; Xiong, Z.; Liang, L. A model transfer strategy based on screening stable wavelength for quantitative analysis of holocellulose and lignin content distribution in pulpwood. *J. Mater. Sci.* **2024**, *59*, 3191–3204. [CrossRef]
16. Xu, Y.X.; Kong, T.Y.; Ma, Y.F.; Zhao, Y.; Chu, L.; Zheng, M.Z. Near-infrared spectroscopy: Application in ensuring food quality and safety. *Anal. Methods* **2025**, *17*, 3381–3406. [CrossRef] [PubMed]
17. Hu, X.C.; Duan, R.Z.; Wang, J.L.; Li, M.C.; Chen, H.; Zhang, J.; Zeng, L.T. Simultaneous detection of cysteine and glutathione in food with a two-channel near-infrared fluorescent probe. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2025**, *325*, 9. [CrossRef]
18. Vera, W.; Salvador-Reyes, R.; Quispe-Santivañez, G.; Kemper, G. Detection of adulterants in powdered foods using near-infrared spectroscopy and chemometrics: Recent advances, challenges, and future perspectives. *Foods* **2025**, *14*, 3195. [CrossRef]
19. Zhou, M.; Liu, C.; Lin, Y.; Bai, T.; Ye, T.; Qian, Z.; Li, L.; Guo, L.; Liu, H.; Wang, J. Novel near-infrared fluorescent probe with large stokes shift for detection of cysteine in food sample and living cells. *Food Chem.* **2025**, *464*, 141755. [CrossRef] [PubMed]

20. Park, R.; Gordon, F.; Agnew, R.; Barnes, R.; Steen, R. The use of near infrared reflectance spectroscopy on dried samples to predict biological parameters of grass silage. *Anim. Feed Sci. Technol.* **1997**, *68*, 235–246. [CrossRef]
21. Rego, G.; Ferrero, F.; Valledor, M.; Campo, J.C.; Forcada, S.; Royo, L.J.; Soldado, A. A portable IoT NIR spectroscopic system to analyze the quality of dairy farm forage. *Comput. Electron. Agric.* **2020**, *175*, 105578. [CrossRef]
22. Shenk, J.; Westerhaus, M. The application of near infrared reflectance spectroscopy (NIRS) to forage analysis. In *Forage Quality, Evaluation, and Utilization*; Amer Society of Agronomy: Madison, WI, USA, 1994; pp. 406–449.
23. Boiret, M.; Chauchard, F. Use of near-infrared spectroscopy and multipoint measurements for quality control of pharmaceutical drug products. *Anal. Bioanal. Chem.* **2017**, *409*, 683–691. [CrossRef] [PubMed]
24. Xu, X.; Xie, L.; Ying, Y. Factors influencing near infrared spectroscopy analysis of agro-products: A review. *Front. Agric. Sci. Eng.* **2019**, *6*, 105–115. [CrossRef]
25. Okura, T. Hardware of near-infrared spectroscopy. In *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 235–264.
26. Zhang, J.; Cheng, Z.; Dong, J.; Zhang, X. Cascaded nanobeam spectrometer with high resolution and scalability. *Optica* **2022**, *9*, 517–521. [CrossRef]
27. Pu, Y.-Y.; O'Donnell, C.; Tobin, J.T.; O'Shea, N. Review of near-infrared spectroscopy as a process analytical technology for real-time product monitoring in dairy processing. *Int. Dairy J.* **2020**, *103*, 104623. [CrossRef]
28. Qi, W.; Jiang, Q.; Erihemu. Research progress on the application of near-infrared spectroscopy in liquid food quality testing. *Beni-Suef Univ. J. Basic Appl. Sci.* **2025**, *14*, 75. [CrossRef]
29. Khodabakhshian, R.; Emadi, B.; Khojastehpour, M.; Golzarian, M.R. A comparative study of reflectance and transmittance modes of Vis/NIR spectroscopy used in determining internal quality attributes in pomegranate fruits. *J. Food Meas. Charact.* **2019**, *13*, 3130–3139. [CrossRef]
30. Hodson, T.O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev. Discuss.* **2022**, *15*, 5481–5487. [CrossRef]
31. Plevris, V.; Solorzano, G.; Bakas, N.P.; Ben Seghier, M.E.A. Investigation of performance metrics in regression analysis and machine learning-based prediction models. In Proceedings of the 8th European Congress on Computational Methods in Applied Sciences and Engineerin, Oslo, Norway, 5–9 June 2022.
32. Cozzolino, D.; Cynkar, W.; Shah, N.; Smith, P. Multivariate data analysis applied to spectroscopy: Potential application to juice and fruit quality. *Food Res. Int.* **2011**, *44*, 1888–1896. [CrossRef]
33. Luo, N.; Han, P.; Wang, S.; Wang, D.; Zhao, C. Near-infrared spectroscopy analytical model using ensemble partial least squares regression. *Anal. Lett.* **2019**, *52*, 1732–1756. [CrossRef]
34. Anjos, O.; Caldeira, I.; Fernandes, T.A.; Pedro, S.I.; Vitória, C.; Oliveira-Alves, S.; Catarino, S.; Canas, S. PLS-R calibration models for wine spirit volatile phenols prediction by near-infrared spectroscopy. *Sensors* **2021**, *22*, 286. [CrossRef]
35. Jin, S.; Sun, F.; Hu, Z.; Li, Y.; Zhao, Z.; Du, G.; Shi, G.; Chen, J. Online quantitative substrate, product, and cell concentration in citric acid fermentation using near-infrared spectroscopy combined with chemometrics. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2023**, *285*, 121842. [CrossRef]
36. Kasemsumran, S.; Boondaeng, A.; Ngowsuwan, K.; Jungtheerapanich, S.; Apiwatanapiwat, W.; Janchai, P.; Meelaksana, J.; Vaithanomsat, P. Simultaneous monitoring of the evolution of chemical parameters in the fermentation process of pineapple fruit wine using the liquid probe for near-infrared coupled with chemometrics. *Foods* **2022**, *11*, 377. [CrossRef]
37. Andrés, J.; Bona, M. Analysis of coal by diffuse reflectance near-infrared spectroscopy. *Anal. Chim. Acta* **2005**, *535*, 123–132. [CrossRef]
38. Liu, C.; Tang, X.; Yu, T.; Wang, T.; Lu, Z.; Yu, W. Measurement of moisture content in lubricating oils of high-speed rail gearbox by Vis-NIR spectroscopy. *Optik* **2020**, *224*, 165694. [CrossRef]
39. Fang, R.; Han, X.; Li, X.; Tong, J.; Gao, M.; Wang, Y. Variable selection in near-infrared spectra for modeling of hemoglobin content in bio-water solutions. *J. Appl. Spectrosc.* **2024**, *91*, 928–935. [CrossRef]
40. Hong, F.W.; Chia, K.S. A review on recent near infrared spectroscopic measurement setups and their challenges. *Measurement* **2021**, *171*, 108732. [CrossRef]
41. Di Egidio, V.; Sinelli, N.; Giovanelli, G.; Moles, A.; Casiraghi, E. NIR and MIR spectroscopy as rapid methods to monitor red wine fermentation. *Eur. Food Res. Technol.* **2010**, *230*, 947–955. [CrossRef]
42. Jeon, K.J.; Hwang, I.D.; Hahn, S.; Yoon, G. Comparison between transmittance and reflectance measurements in glucose determination using near infrared spectroscopy. *J. Biomed. Opt.* **2006**, *11*, 014022. [CrossRef]
43. Yang, M.; Dhanabalan, S.S.; Robel, M.R.; Thekkekara, L.V.; Mahasivam, S.; Rahman, M.A.; Borkhatariya, S.; Sen, S.; Walia, S.; Sriram, S. Miniaturized optical glucose sensor using 1600–1700 nm near-infrared light. *Adv. Sens. Res.* **2025**, *4*, 2300160. [CrossRef]

44. Siesler, H.W.; Ozaki, Y.; Kawata, S.; Heise, H.M. *Near-Infrared Spectroscopy: Principles, Instruments, Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
45. Rinnan, Å.; Van Den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222. [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.