Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Development of Symbolic Regression-Based ATR-FTIR Calibration Models

Fernando A. R. D. Lima[abc]*, Inga S. Nordhus[a], Marcellus G. F. de Moraes[c], M. Enis Leblebici[d], Argimiro R. Secchi[bc], Mauricio B. de Souza Jr.[bc] and Idelfonso Nogueira[a]

[a] Chemical Engineering Department, Norwegian University of Science and Technology, Trondheim, 793101, Norway
[b] EPQB, School of Chemistry, Universidade Federal do Rio de Janeiro, Av. Horácio Macedo, 2030, CT, Bloco E, 21941-914, Rio de Janeiro, RJ – Brazil
[c] Chemical Engineering Program, PEQ/COPPE – Universidade Federal do Rio de Janeiro, Av. Horácio Macedo, 2030, CT, Bloco G, G115, 21941-914, Rio de Janeiro, RJ – Brazil
[d] KU Leuven, Center for Industrial Process Technology, Diepenbeek, Belgium
* Corresponding Author: farrais@eq.ufrj.br.

ABSTRACT

Accurate calibration of spectroscopic measurements is essential for reliable real-time monitoring and control of crystallization processes. In this work, calibration strategies for Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) spectroscopy were systematically evaluated for concentration monitoring in batch cooling crystallization of paracetamol in ethanol. Linear regression (LR), Partial Least Squares Regression (PLSR), Principal Component Regression (PCR), and symbolic regression (SR) were compared using both peak-based features and full spectral representations. Peak-based models provided a transparent baseline, with peak-area-based models consistently outperforming peak-height-based models. For LR, incorporating multiple absorption bands reduced the mean squared error (MSE) by nearly one order of magnitude compared to single-peak models. Using the same peak-based inputs, SR further improved performance, reducing prediction bias at high concentrations and yielding higher coefficients of determination (R^2 > 0.99) compared to LR. A substantial improvement was achieved when full spectral information was used. Among all evaluated approaches, SR with unprocessed spectra yielded the best overall performance, achieving an R^2 of 0.996 and an MSE of 1.4×10^{-6} on the validation dataset. This model also demonstrated strong generalization on an independent solubility test dataset, closely reproducing the reference solubility curve over the full temperature range with minimal deviation. In contrast, PCR and PLSR models showed increased sensitivity to preprocessing choices and exhibited larger errors on the test dataset. SR provided an accurate, robust, and interpretable calibration framework for ATR-FTIR, with reduced reliance on spectral preprocessing and potential for real-time process analytical technology and control applications.

Keywords: PAT, Preprocessing, Process monitoring, PLSR, Crystallization

INTRODUCTION

Crystallization is a key separation and purification unit operation in the pharmaceutical and fine chemicals industries, where stringent quality requirements demand tight control over product attributes such as purity, crystal size distribution (CSD), polymorphic form, and morphology [1]. In batch cooling crystallization, these attributes are influenced by the temporal evolution of supersaturation, which governs nucleation and crystal growth kinetics. Poor supersaturation control can lead to excessive primary or secondary nucleation, agglomeration, fouling, and batch-to-batch variability, ultimately compromising product quality and process robustness [2]. As a result, real-time monitoring and control of crystallization processes have been a long-standing focus of research in process systems engineering and control [3].

The introduction of process analytical technology (PAT) frameworks has significantly advanced the monitoring and understanding of crystallization systems [4].

PAT tools enable real-time or near-real-time access to critical process information, supporting data-driven decision-making and advanced control strategies [5]. Commonly applied PAT techniques in crystallization include attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy, Raman spectroscopy, focused beam reflectance measurement (FBRM), particle vision and measurement (PVM), and ultrasonic spectroscopy [6]. Spectroscopic techniques such as ATR-FTIR and Raman are primarily used for monitoring solute concentration and supersaturation, whereas chord length distribution (CLD) measurements from FBRM and image-based measurements from PVM provide indirect information related to particle size and shape. The complementary use of multiple PAT tools has enabled more comprehensive characterization of crystallization dynamics and facilitated the development of model-based monitoring and control frameworks [7].

Among these techniques, ATR-FTIR spectroscopy has become one of the most widely adopted PAT tools for concentration measurement in crystallization processes [8]. Its popularity stems from its non-invasive operation, fast sampling rates, robustness in slurry environments, and compatibility with laboratory and pilot-scale crystallizers. However, ATR-FTIR does not provide direct concentration measurements; instead, it generates spectra that must be translated into quantitative concentration estimates through calibration models [7]. The reliability of any ATR-FTIR-based monitoring or control strategy therefore depends fundamentally on the accuracy, robustness, and generalization capability of the calibration model [4, 6].

Early ATR-FTIR calibration approaches were predominantly based on linear regression (LR) models using selected peak heights or peak areas, often combined with baseline correction and explicit temperature compensation [9]. While these methods are straightforward to implement and interpret, they rely heavily on expert-driven feature selection and may fail to capture nonlinearities arising from temperature effects, solute–solvent interactions, and spectral overlap. To address these limitations, multivariate statistical techniques such as partial least squares regression (PLSR) have become the standard approach to develop ATR-FTIR calibration models [6]. PLSR enables dimensionality reduction while maximizing covariance between spectral data and concentration and has been successfully applied in a wide range of crystallization studies [10, 11].

Despite their widespread use, conventional calibration approaches exhibit several limitations. Linear models can be overly restrictive, while PLSR models are often sensitive to data preprocessing choices, including spectral filtering, scatter correction, and wavelength range selection [12]. Furthermore, latent-variable models such as PLSR may suffer from reduced interpretability, as the physical meaning of latent components is not always clear. These challenges become particularly pronounced when calibration models are expected to operate across varying temperatures, solvent compositions, or process conditions, motivating the exploration of more flexible and robust modeling strategies [4, 6].

In recent years, machine learning techniques have been increasingly investigated as alternatives or complements to classical calibration methods in spectroscopic PAT applications [13]. Approaches such as support vector regression, Gaussian process regression, and artificial neural networks (ANNs) have demonstrated improved predictive performance in some cases, particularly when nonlinear effects are significant [6]. However, many of these methods function as black-box models, limiting their interpretability and complicating their deployment in regulated industrial environments [4]. Moreover, their computational complexity and reliance on large, representative datasets can pose challenges for real-time implementation and long-term maintenance.

Within this context, symbolic regression (SR) has emerged as a promising interpretable machine learning approach for process modeling and calibration [7]. SR aims to identify explicit mathematical expressions that relate input variables to outputs, without assuming a predefined model structure [14]. By balancing model accuracy and complexity, SR can uncover nonlinear relationships while producing closed-form equations that are transparent and amenable to physical interpretation [15]. These characteristics make SR particularly attractive for PAT-enabled crystallization systems, where interpretability, robustness, and ease of deployment are critical.

Although SR has been successfully applied to a variety of chemical engineering problems, including model correction, system identification, and control-oriented modeling [14-16], its application to ATR-FTIR calibration for crystallization monitoring remains limited in the open literature [7]. Most existing studies continue to rely on LR or latent-variable methods, despite growing interest in interpretable scientific machine learning frameworks [4, 6]. This gap highlights the need for a systematic evaluation of SR as a calibration methodology and a direct comparison with established approaches under consistent experimental conditions.

In this work, ATR-FTIR calibration models were developed for paracetamol–ethanol systems using SR and benchmarked against conventional LR, PLSR, and and principal component regression (PCR) models. As presented in Figure 1, different input representations were investigated, including peak-based features, temperature, and principal components derived from limited spectral ranges. The influence of common preprocessing strategies was assessed, and the resulting models were evaluated in terms of predictive accuracy, robustness, and generalization capability. The main objective was to

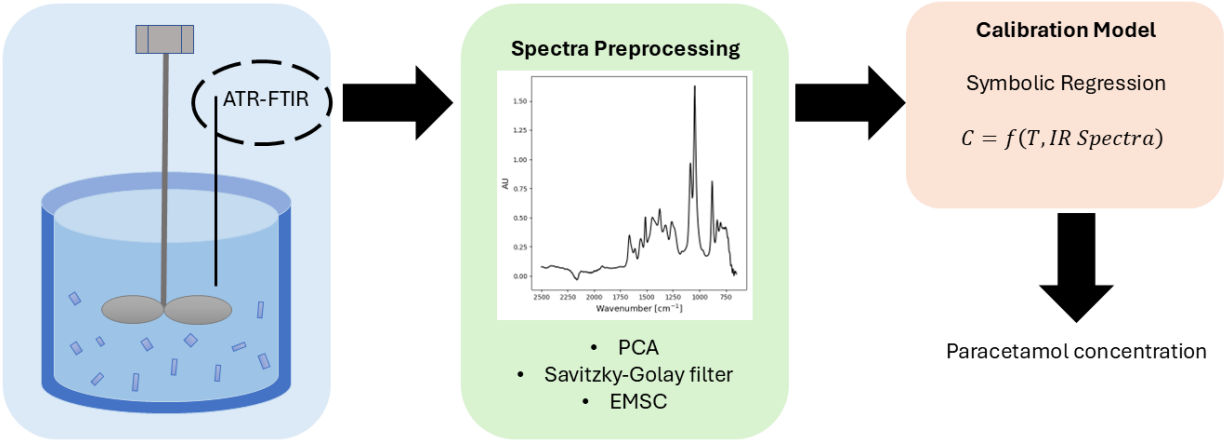

Figure 1. Symbolic regression approach proposed to develop ATR-FTIR calibration models.

assess the suitability of SR as a practical and interpretable alternative for ATR-FTIR calibration in crystallization monitoring and control applications.

DATA ACQUISITION

The experimental dataset used in this work was obtained from the batch crystallization study reported in Lima et al. [7]. In that work, paracetamol–ethanol mixtures were prepared gravimetrically and monitored in situ using ATR-FTIR spectroscopy, with complementary PAT tools (FBRM and temperature sensing) available in the same setup. Infrared spectra were acquired using a ReactIR system over 648–2500 cm^{-1} at a sampling rate of one spectrum per minute.

For calibration-model development, the dataset comprises 107 ATR-FTIR spectra paired with reference concentrations covering 0 to 0.5 g paracetamol/g ethanol and temperatures spanning from 10 to 70 °C. The data originate from four calibration experiments designed to systematically vary composition and temperature, including dilution sequences where ethanol was added to a fully dissolved paracetamol–ethanol solution and stepwise addition of paracetamol into ethanol. Across these experiments, spectra for a given concentration level were acquired at multiple temperatures, ensuring that the calibration models capture the coupled influence of concentration and temperature on spectral response. This dataset was randomly divided into calibration and validation subsets, with 70 % of the samples used for model training and the remaining 30 % reserved for validation.

In addition to the calibration and validation datasets, an independent test dataset was used to evaluate model generalization. This test set corresponds to an equilibrium solubility experiment reported in Lima et al. [7], where a paracetamol–ethanol mixture was held at 10 °C for 3 h to ensure equilibrium and then heated at 0.1 K/min up to 55 °C. The resulting equilibrium concentrations were compared against the solubility model reported by Griffin et al. [17], providing an external check that model-predicted concentrations reproduce physically meaningful temperature–solubility trends beyond the calibration split.

SPECTRAL PREPROCESSING AND FEATURE CONSTRUCTION

Preprocessing approaches were systematically tested to quantify how common chemometric treatments and feature representations affect calibration accuracy and robustness. The starting point was a peak-based calibration strategy derived from chemically meaningful absorption bands of paracetamol in the ATR-FTIR spectra. Both peak height and peak area were evaluated as inputs. Prior to peak extraction, baseline subtraction was applied to reduce offsets and slow drifts in the spectra. Peak areas and peak heights were then computed over the following wavenumber intervals: 1584-1532 cm^{-1}, 1532-1492 cm^{-1}, and 1292-1184 cm^{-1}.

Beyond the peak-based representation, full-spectrum representations were also investigated to evaluate whether multivariate inputs improve performance relative to hand-crafted features. In addition to the full spectral range, a reduced window between 600 and 1800 cm^{-1} was considered, as it concentrates the most informative absorbance features for paracetamol while excluding regions more affected by noise or solvent interference. For noise attenuation, Savitzky–Golay filtering was tested using SciPy [18]. To compensate for baseline shifts and multiplicative scattering effects, extended multiplicative signal correction (EMSC) was evaluated using the chemometrics Python package [19], both alone and combined with smoothing. When spectral vectors (full or reduced) were used as model inputs, dimensionality reduction via principal component analysis (PCA) was performed using scikit-learn [20], and the resulting

components were used as regressors together with temperature.

All preprocessing variants were applied consistently across training, validation, and test datasets, and preprocessing parameters were kept fixed across modeling approaches to avoid information leakage and ensure that observed performance differences reflect the calibration models rather than data handling choices.

CALIBRATION MODEL DEVELOPMENT

Four calibration strategies were investigated to calculate the paracetamol mole fraction and compared in this work: LR, PLSR, PCR, and SR. LR models were constructed using peak-based spectral features obtained after baseline subtraction. Both peak heights and peak areas were considered as explanatory variables, together with temperature to account for its influence on spectral response and solubility. Model parameters were estimated using ordinary least squares as implemented in scikit-learn [20]. These linear models were directly compared to SR models using the same peak-area and peak-height inputs, enabling an explicit assessment of whether SR provides benefit over the conventional peak-based calibration strategy.

PLSR and PCR models were developed using spectral information in combination with temperature, using implementations available in scikit-learn [20]. Depending on the preprocessing strategy, either full-spectrum data or reduced spectral windows were used as inputs. For PCR, PCA was first applied to the spectral matrix, and a fixed number of principal components were retained as regressors in a linear model together with temperature. For PLSR, latent variables were extracted by maximizing covariance between spectral inputs and concentration. In both PCR and PLSR, the number of retained components/latent variables was selected a priori by inspecting the cumulative explained variance of the spectral decomposition and the stability of prediction performance on the held-out validation split, and then fixing a single value used consistently across all subsequent comparisons.

SR was employed as an interpretable machine learning approach to develop nonlinear calibration models without assuming a predefined functional form. SR models were developed using the PySR framework [21], which searches for explicit mathematical expressions relating spectral inputs and temperature to solute concentration. Different input representations were explored, including peak-based features (peak height/area) and principal components derived from full or reduced spectra. Model complexity was controlled by limiting expression depth and length to promote models suitable for real-time monitoring and control applications.

Across all calibration strategies, model performance was evaluated using identical data splits and performance metrics, ensuring a consistent and fair comparison between linear, latent-variable, and symbolic approaches.

RESULTS AND DISCUSSION

Peak-based Calibration Models

Initially, calibration models based on manually selected spectral features were investigated. Peak-based models were constructed using both peak height and peak area extracted from characteristic paracetamol absorption bands after baseline subtraction. The models were initially developed using a single absorption band and subsequently extended by including additional chemically relevant peaks, specifically the 1292–1184 cm^{-1}, 1532–1492 cm^{-1}, and 1584–1532 cm^{-1} regions. These models represent the conventional approach adopted in ATR-FTIR calibration and provide a benchmark against which multivariate and symbolic regression models can be assessed. The SR models were trained with 200 iterations and a population size of 20.

Table 1: Performance of the LR and SR models to calculate the paracetamol mole fraction for the validation set, considering temperature and peak areas or heights as inputs

Peak Area			
Model	**Number of Inputs**	**R^2**	**MSE**
LR	2	0.982	1.40×10^{-5}
LR	3	0.984	1.31×10^{-5}
LR	4	0.990	7.74×10^{-6}
SR	2	0.985	1.21×10^{-5}
SR	3	0.991	7.16×10^{-6}
SR	4	0.994	4.41×10^{-6}
Peak Height			
Model	**Number of Inputs**	**R^2**	**MSE**
LR	2	0.986	1.08×10^{-5}
LR	3	0.986	1.14×10^{-5}
LR	4	0.993	5.75×10^{-6}
SR	2	0.976	3.49×10^{-3}
SR	3	0.992	5.97×10^{-6}
SR	4	0.997	2.37×10^{-6}

Table 1 summarizes the performance of the LR and SR models developed using peak-based spectral features and temperature as inputs. For both peak height and peak area representations, increasing the number of peaks used as inputs leads to a systematic improvement in model accuracy, as reflected by increasing R^2 values and decreasing mean squared error (MSE). This trend is observed consistently for both LR and SR models and indicates that incorporating multiple chemically relevant

absorption bands provides complementary information, reducing sensitivity to noise and local spectral distortions associated with individual peaks.

For LR models, the improvement with additional peaks is particularly evident when moving from two to four inputs, where the MSE decreases significantly for both peak height and peak area representations. This reflects the limited expressive capability of linear models, which rely on sufficiently rich input information to approximate the underlying concentration–temperature relationship. Eq. 1 shows the resulting LR formulation, where the predicted concentration is expressed as a weighted linear combination of peak intensities and temperature. While this structure is transparent and easy to interpret, it cannot explicitly represent nonlinear interactions between spectral features and temperature.

$$X_{para} = -0.000191T + 0.271\,h_{1256} + 0.356h_{1512} - 0.559h_{1560} + 0.0447 \quad (1)$$

On the other hand, the SR model, given in Eq. 2, introduces nonlinear combinations of peak intensities and temperature through higher-order and multiplicative terms. These nonlinear structures allow the model to capture temperature-dependent changes in spectral response and interactions between absorption bands that are not accessible to linear regression. As a result, SR achieves lower MSE and higher R^2 values than LR for the same set of peak-based inputs, particularly when three or four peaks are included. The model described in Eq. 2 was obtained using the three peak heights as inputs, but only the most relevant inputs stay in the equation in symbolic regression.

$$X_{para} = h_{1512}[\,(0.714 + 0.430h_{1512})^4 - 0.00553h_{1560}T] \quad (2)$$

Figure 2 presents the parity plots for LR and SR models using four peak heights and temperature as inputs. Both models show good agreement between predicted and reference concentrations. However, the SR model exhibited reduced bias at higher concentrations and improved clustering around the parity line. This behavior was consistent with the nonlinear structure of Eq. 2, which enables better extrapolation toward the solubility limit compared to the linear formulation in Eq. 1.

Despite these improvements, peak-based models remain fundamentally limited by the reduced information content of manually selected spectral features. Although SR partially mitigates this limitation through nonlinear functional forms, the performance gains remain modest when compared to multivariate models based on full or reduced spectral representations. Nevertheless, the peak-based analysis establishes a clear and interpretable baseline, demonstrating that SR can extract additional predictive value from low-dimensional inputs while preserving model transparency and suitability for real-time implementation.

Spectra-based Calibration Models

To overcome the limitations associated with manually selected peak-based features, calibration models based on spectral representations were subsequently investigated. In this approach, the ATR-FTIR spectra were used directly as model inputs together with temperature, allowing the calibration models to exploit distributed spectral information related to both concentration and temperature effects. The comparative performance of the spectra-based calibration models is summarized in Table 2 for the validation set, while representative parity plots for the validation dataset are shown in Fig. 3.

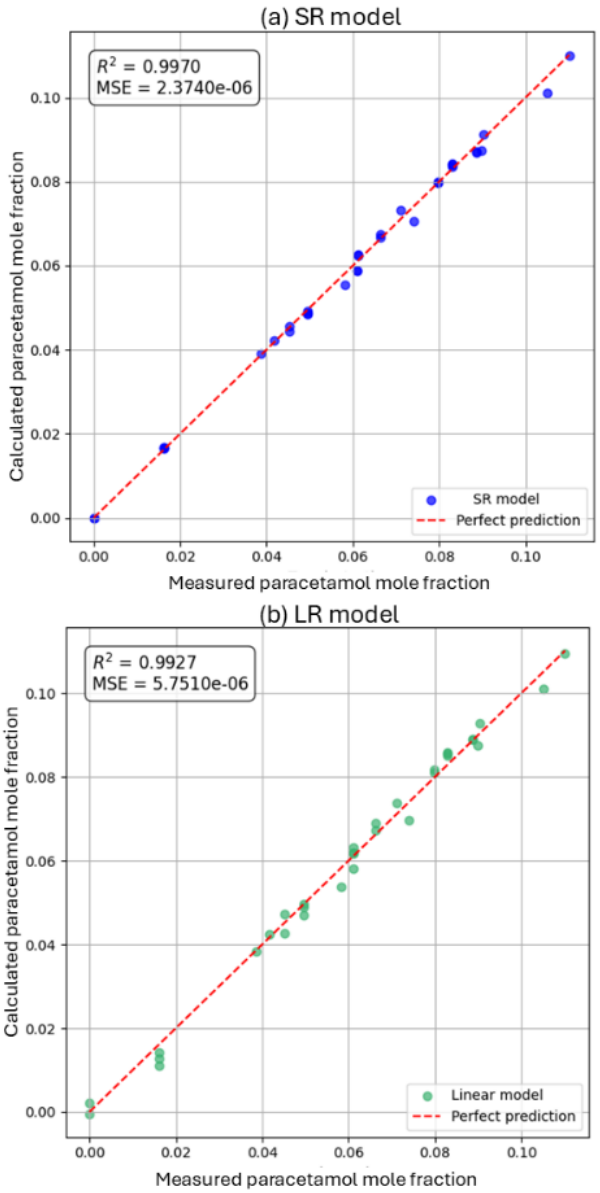

Figure 2. Comparison between the predicted and measured values of the paracetamol mole fraction for the validation set, considering SR and LR models using four peak heights and temperature as input.

Table 2: Performance of spectra-based calibration models considering different preprocessing strategies for the validation set

	Metric	Full spectra			Trimmed spectra		
		SR	PLSR	PCR	SR	PLSR	PCR
No preprocessing	R^2	0.996	0.974	0.971	0.992	0.974	0.9708
	MSE	1.40×10^{-6}	9.83×10^{-6}	1.12×10^{-5}	3.18×10^{-6}	9.83×10^{-6}	1.12×10^{-5}
SG filter	R^2	0.997	0.998	0.997	0.998	0.998	0.997
	MSE	1.15×10^{-6}	8.48×10^{-7}	1.03×10^{-6}	8.06×10^{-7}	9.60×10^{-7}	1.20×10^{-6}
EMSC	R^2	0.994	0.994	0.992	0.992	0.994	0.992
	MSE	2.31×10^{-6}	2.32×10^{-6}	2.91×10^{-6}	3.10×10^{-6}	2.32×10^{-6}	2.91×10^{-6}
EMSC + SG filter	R^2	0.995	0.997	0.996	0.992	0.992	0.996
	MSE	1.83×10^{-6}	1.03×10^{-6}	1.62×10^{-6}	3.11×10^{-6}	2.91×10^{-6}	1.48×10^{-6}

The influence of spectral preprocessing was systematically evaluated using common chemometric approaches, including Savitzky–Golay (SG) filtering, EMSC, and spectral range reduction. As presented in Table 2, these preprocessing steps noticeably affect linear and latent-variable models, but have a limited impact on SR. Notably, the best overall performance was obtained with SR trained on the full, unprocessed spectra, indicating that SR can internally construct nonlinear combinations of spectral features that effectively account for baseline variations, noise, and temperature-dependent distortions without requiring explicit preprocessing. In several cases, applying preprocessing led to negligible improvements or slight degradations, reinforcing that extensive preprocessing is not necessary for the SR workflow in this dataset.

As shown in Table 2, spectra-based models consistently outperformed peak-based strategies across all modeling approaches. This improvement is particularly pronounced for SR, which achieves the lowest prediction errors and highest coefficients of determination. The parity plots in Fig. 3 presents the improved agreement between predicted and measured concentrations when full spectral information is used, with reduced scatter and minimal systematic bias across the calibration and validation datasets.

The calibration model obtained by SR using the raw spectra and temperature as inputs is described in Eq. 3. $\hat{h}_i$ is the absorbance intensity at wavenumber i cm^{-1}. All SR models were trained with 200 iterations and a population size of 20.

$$X_{para} = \hat{h}_{1656}(1.148 - \hat{h}_{1056}) + 2.425(\hat{h}_{2128}^2 - 0.00419) \quad (3)$$

Eq. 3 highlights the ability of SR to identify compact and interpretable nonlinear relationships directly from raw spectral inputs. The model combines contributions from multiple spectral regions through both linear and nonlinear terms, capturing interactions that are not explicitly represented in linear or latent-variable models.

The generalization capability of the spectra-based models was further assessed using an independent test dataset, corresponding to equilibrium solubility measurements. The performance on this test dataset is shown in Fig. 4, in which model predictions are compared against reference solubility model as a function of temperature from Griffin et al. [17]. The SR model trained on the full, unprocessed spectra shows close agreement with the reference solubility behavior over the investigated temperature range, with only minor deviations near the solubility limit. The performance of the SR model confirms the extrapolation capability beyond the calibration split and demonstrates that the model preserves physically meaningful concentration–temperature trends under equilibrium conditions.

PCR and PLSR models exhibited more dependence on preprocessing choices. Although dimensionality reduction and scatter correction can improve calibration and validation metrics, increased deviation from the reference solubility trend was observed on the test dataset, as shown in Fig. 4, indicating reduced robustness when

extrapolating toward equilibrium conditions. Overall, the spectra-based analysis demonstrates that combining full ATR-FTIR spectra with SR yielded the most accurate and robust calibration strategy investigated in this study, and that its accurate performance on the independent solubility test dataset supports its suitability for reliable spectroscopic monitoring and control of crystallization processes.

CONCLUSION

This study evaluated different calibration strategies for ATR-FTIR-based concentration monitoring, comparing conventional peak-based models with spectra-based approaches using LR, latent-variable methods, and SR. Peak-based models provided a transparent baseline, with improved performance when multiple absorption bands and peak areas were used. SR consistently outperformed LR for the same peak-based inputs by capturing nonlinear interactions between spectral features and temperature, although performance remained limited by the reduced information content of manually selected peaks.

The use of full spectral information led to a substantial improvement in calibration accuracy and robustness. Among all approaches, SR applied directly to the unprocessed spectra yielded the best performance and demonstrated generalization on an independent solubility test dataset. The ability of SR to deliver compact, interpretable models without extensive preprocessing highlights its potential as a practical and robust calibration methodology for real-time spectroscopic monitoring and control of crystallization processes.

ACKNOWLEDGEMENTS

This study was financed in part by CAPES - Finance Code 001 and by Petrobras S.A. (Cooperation term no. 0050.0125244.23.9). Professors MB de Souza Jr. and AR Secchi are grateful to financial support from CNPq (Grants No. 311153/2021-6 and 300744/2025-0).

AUTHOR IDENTIFIERS

Author ORCIDs:
Lima FARD: 0009-0004-5025-3804
Nordhus IS: -
Moraes MGF: 0000-0003-1580-3669
Leblebici ME: 0000-0003-4599-9412
Secchi AR: 0000-0001-7297-3571
Souza MB Jr: 0000-0002-1090-8958
Nogueira IBR: 0000-0002-0963-6449

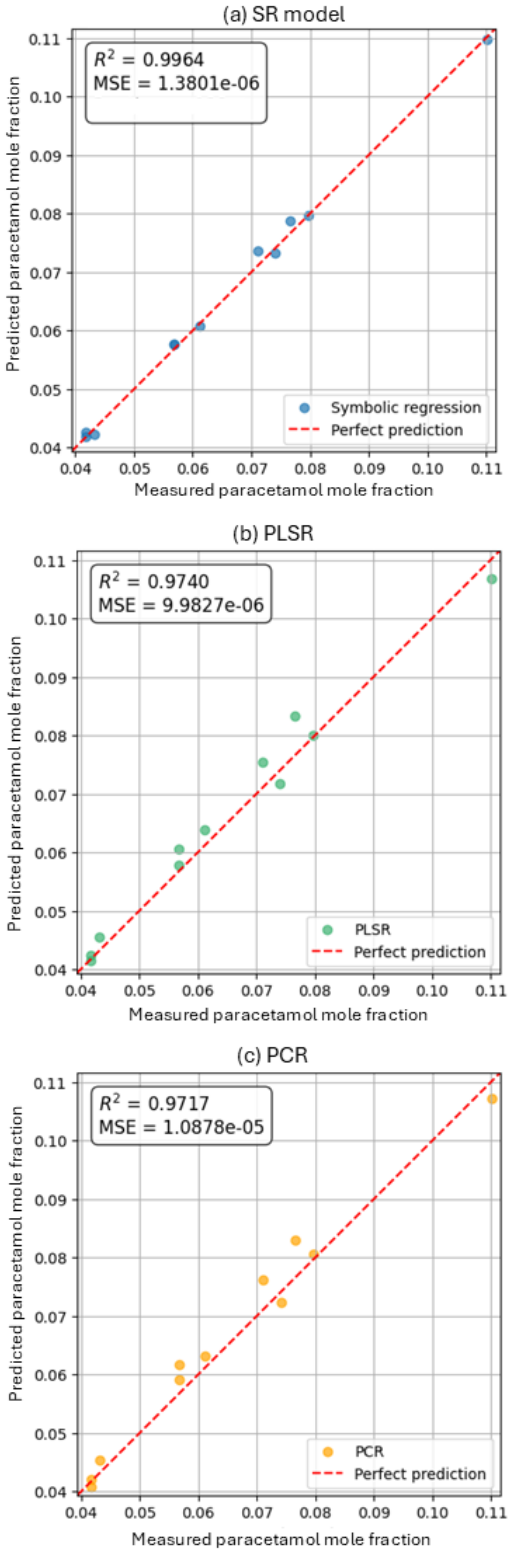

Figure 3. Comparison between the predicted and measured values of the paracetamol mole fraction for the validation set, considering SR, PLSR and PCR models using all spectra and temperature as input.

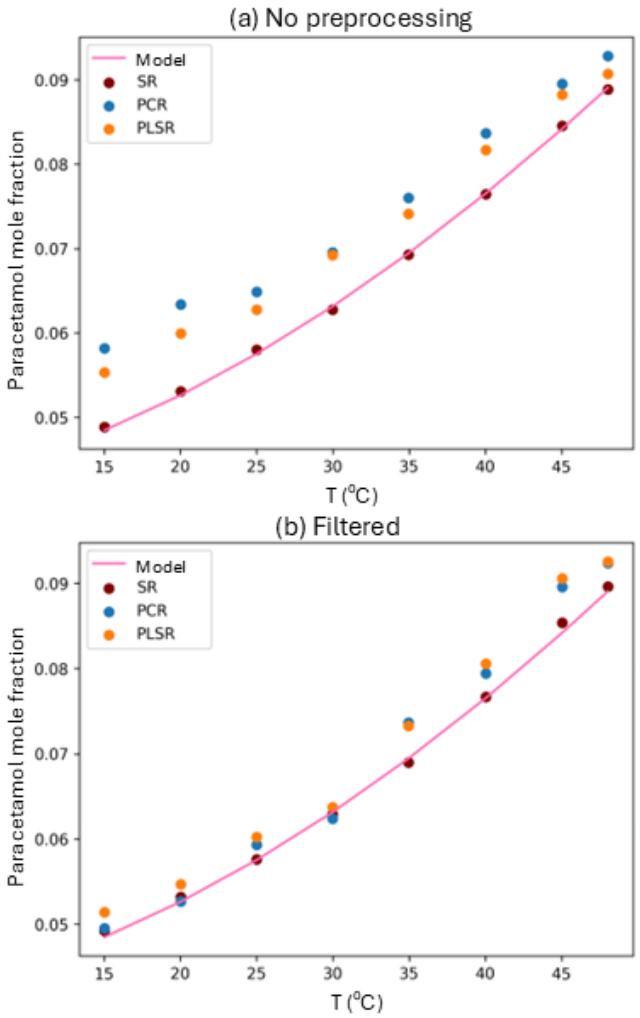

Figure 4. Comparison between the predictions of SR, PLSR and PCR to the solubility model of Griffin et al. [17], considering no preprocessing and filtering.

REFERENCES

1. Gao Z, Rohani S, Gong J, Wang J. Recent developments in the crystallization process: toward the pharmaceutical industry. Engineering 3:343-353 (2017). https://doi.org/10.1016/j.eng.2017.03.022
2. Braatz RD. Advanced control of crystallization processes. Annual Reviews in Control 26:87-99 (2002). https://doi.org/10.1016/s1367-5788(02)80016-5
3. Nagy ZK, Braatz RD. Advances and new directions in crystallization control. Annu. Rev. Chem. Biomol. Eng. 3:55-75 (2012). https://doi.org/10.1146/annurev-chembioeng-062011-081043
4. Lima FARD, de Moraes MGF, Barreto AG Jr, Secchi AR, Grover MA, de Souza MB Jr. Applications of machine learning for modeling and advanced control of crystallization processes: developments and perspectives. Digital Chemical Engineering 14:100208 (2025). https://doi.org/10.1016/j.dche.2024.100208
5. Simon LL, et al. Assessment of recent process analytical technology (PAT) trends: a multiauthor review. *Org Process Res Dev* 19:3–62 (2015) https://doi.org/10.1021/op500261y
6. Xiouras C, Cameli F, Quilló GL, Kavousanakis ME, Vlachos DG, Stefanidis GD. Applications of artificial intelligence and machine learning algorithms to crystallization. Chem. Rev. 122:13006-13042 (2022). https://doi.org/10.1021/acs.chemrev.2c00141
7. Dias Lima FAR, Fernandes de Moraes MG, Resende Secchi A, de Souza MB Jr, Grover MA. Experimental nonlinear model predictive control of crystal size and yield in batch cooling crystallization enabled by soft sensor and symbolic-based calibration model. Ind. Eng. Chem. Res. 64:23582-23600 (2025). https://doi.org/10.1021/acs.iecr.5c03894
8. Zhang F, Du K, Guo L, Huo Y, He K, Shan B. Progress, problems, and potential of technology for measuring solution concentration in crystallization processes. Measurement 187:110328 (2022). https://doi.org/10.1016/j.measurement.2021.110328
9. Swinehart DF. The beer-lambert law. J. Chem. Educ. 39:333 (1962). https://doi.org/10.1021/ed039p333
10. Lindenberg C, Krättli M, Cornel J, Mazzotti M, Brozio J. Design and optimization of a combined cooling/antisolvent crystallization process. Crystal Growth & Design 9:1124-1136 (2008). https://doi.org/10.1021/cg800934h
11. Trampuž M, Tesli? D, Likozar B. Process analytical technology-based (PAT) model simulations of a combined cooling, seeded and antisolvent crystallization of an active pharmaceutical ingredient (API). Powder Technology 366:873-890 (2020). https://doi.org/10.1016/j.powtec.2020.03.027
12. Zhang F, Liu T, Wang XZ, Liu J, Jiang X. Comparative study on ATR-FTIR calibration models for monitoring solution concentration in cooling crystallization. Journal of Crystal Growth 459:50-55 (2017). https://doi.org/10.1016/j.jcrysgro.2016.11.064
13. Lu M, Rao S, Yue H, Han J, Wang J. Recent advances in the application of machine learning to crystal behavior and crystallization process control. Crystal Growth & Design 24:5374-5396 (2024). https://doi.org/10.1021/acs.cgd.3c01251
14. Lima FARD, de Moraes MGF, Rebello CM, Barreto AG Jr, Secchi AR, de Souza MB Jr, Nogueira IBR.

Interpretable and uncertainty-aware machine learning for trustworthy prediction in batch crystallization. Chemical Engineering and Processing - Process Intensification 215:110350 (2025). https://doi.org/10.1016/j.cep.2025.110350
15. Rebello CM, Costa EA, Fontana M, Schnitman L, Nogueira IBR. Interpretable scientific machine learning approach for correcting phenomenological models: methodology validation on an ESP prototype. Ind. Eng. Chem. Res. 63:19030-19050 (2024). https://doi.org/10.1021/acs.iecr.4c02104
16. Santana VV, Costa E, Rebello CM, Ribeiro AM, Rackauckas C, Nogueira IBR. Efficient hybrid modeling and sorption model discovery for non-linear advection-diffusion-sorption systems: a systematic scientific machine learning approach. Chemical Engineering Science 282:119223 (2023). https://doi.org/10.1016/j.ces.2023.119223
17. Griffin DJ, Kawajiri Y, Rousseau RW, Grover MA. Using MC plots for control of paracetamol crystallization. Chemical Engineering Science 164:344-360 (2017). https://doi.org/10.1016/j.ces.2017.01.065
18. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17:261–272 (2020) https://doi.org/10.1038/s41592-019-0686-2
19. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems 118:62-69 (2012). https://doi.org/10.1016/j.chemolab.2012.07.010
20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830 (2011)
21. Cranmer K. Interpretable machine learning for science with PySR and symbolic regression. *Mach Learn Sci Technol* 4:015018 (2023) https://doi.org/10.1088/2632-2153/ac9f7c