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

Addressing Matrix Effects Through A Physical Prior-Informed Calibration Model For Quantitative Analysis

Onur C. Boy[a], Ulderico Di Caprio[a], Idelfonso Nogueira[b] and M. Enis Leblebici[a]*

[a] Center for Industrial Process Technology, Department of Chemical Engineering, KU Leuven, Agoralaan Building B, Diepenbeek, Belgium
[b] Department of Chemical Engineering, Norwegian University of Science and Technology, Gløshaugen, Trondheim, Norway
* Corresponding Author: muminenis.leblebici@kuleuven.be

ABSTRACT

Building a robust calibration curve is essential for accurate quantification of multicomponent mixtures. Matrix effects can distort the proportional relationship between the analyte concentration and instrumental response. In addition, classical machine learning models do not inherently incorporate simple physical constraints, such as the requirement that a zero response must correspond to a zero concentration, which can result in non-zero predictions. To address this limitations, prior-induced calibration models were proposed that inherently embed this physical constraint into the model architecture. A dataset was generated for ethanol electrooxidation products using headspace gas-chromatography-mass spectrometry (HS-GC-MS). Multiple linear regression (MLR), polynomial regression and artificial neural network (ANN) models were trained to investigate the effects of model complexity and the incorporation of physical information on predictive performance. Model selection and complexity were assessed using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The prior-induced polynomial regression model achieved improved predictive accuracy while maintaining low model complexity, whereas the ANN provided comparable accuracy but was strongly penalized due to its substantially higher complexity.

Keywords: Matrix effects, Calibration curve, Polynomial regression, Artificial neural network, Ethanol electrooxidation, Chemometrics

INTRODUCTION

Accurate quantification of individual species in multicomponent mixtures is essential for reliable chemical analysis, process monitoring and optimization of operating parameters on selectivity and conversion, which is critical for both batch studies and reactor design. Matrix effects and competitive partitioning among analytes complicate accurate quantification by altering the linear relationship between the analyte concentration and the response [1]. In such systems, the measured peak area is not solely proportional to the analyte concentration but is also influenced by the concentrations of the co-existing components, since each composition defines a new vapor–liquid equilibrium governed by non-ideal intermolecular interactions. To overcome this, an appropriate calibration model is required to accurately resolve the concentration of each component [2]. A calibration curve is a regression model between the known concentrations of analytes and its response obtained from an analytical instrument, which allows the prediction of the concentration of an unknown mixture [3].

Matrix effects have been addressed using calibration models, including linear and non-linear models such as MLR, principal component analysis, partial least squares (PLS), and ANNs [4]. Veiga-del-Baño et al. built LC–MS/MS and GC–MS/MS calibration models using linear, weighted linear, and polynomial regression, and reported that the weighted linear model had the best performance compared the other models [5]. Also, to identify adequate calibration models, Chen et al. investigated six different linear and non-linear calibration equations and reported that most real calibration curves exhibit measurable non-linearity, with non-linear models often outperforming linear models in terms of fitting and prediction [6]. Matrix effects were addressed by developing a calibration curve using PLS model to quantify volatile compounds in wine, where single internal standard

calibration failed to correct matrix effects [7]. Campmajó et al. used PLS to quantify peanut and hazelnut adulteration in almond products with prediction errors below 6.1% and no observable matrix effects [8]. Brusamarello et al. compared PLS and ANN models on UV–Vis spectra for soybean biodiesel quantification, finding that the ANN model achieved higher predictive accuracy for complex reaction mixtures [9]. Gholivand et al. addressed strong serum matrix effects in voltammetric analysis by calibrating PLS models in blank serum and applying baseline and alignment corrections, allowing reliable simultaneous determination of five electroactive alkaloids [10].

Classical machine learning algorithms cannot inherently incorporate simple physical constraints, such as the condition that in the absence of a given species, the corresponding peak area should be zero. As a result, such models may produce non-zero predictions even when the input peak area is zero. This issue can be addressed by applying a rule-based correction layer that forces the output to zero whenever the input value is zero or by manipulating the dataset by augmenting the training dataset with zero-concentration samples, however, this could introduce bias into the model. Alternatively, the model can be designed to inherently learn this constraint. In this case, the model cannot violate the physical boundary condition, as the physical information is enforced by design. Moreover, embedding this constraint reduces the effective model complexity and improves robustness near the detection limit.

This work aims to develop an data-driven calibration model for quantification of ethanol oxidation products while capturing the nonlinear relationships arising from matrix effects. The model also incorporates a simple physical prior: the principle that a zero peak area corresponds to zero concentration to increase prediction accuracy further. There are two models proposed: Prior-informed polynomial regression and prior-informed branched ANN model. The predictive performances of these models are compared with those of MLR, polynomial regression, and feed forward ANN. The model complexities are assessed using AIC and BIC.

METHODOLOGY

Chemicals

Ethanol (>99%) was purchased from Fisher Scientific (Hampton, NH, USA). Acetaldehyde (>99.5%), acetic acid (>99.9%), ethyl acetate (>99.7%), and 1-propanol (>99.9%) were obtained from Sigma-Aldrich (St. Louis, MO, USA). All chemicals were of analytical grade and used as received without further purification. Ultrapure water (resistivity ≥18.2 MΩ·cm) was used for all solution preparations

Method

Headspace gas chromatography–mass spectrometry (HS–GC–MS) was utilized to generate a dataset of standard solutions. Analyses were performed using an Agilent 8890 GC coupled to a 5977C MS and an 8697 HS. Chromatographic separation was achieved using an Agilent J&W HP-5 capillary column (30 m × 0.25 mm × 0.50 µm). Helium was used as the carrier gas at a constant flow rate of 2 mL min^{-1}. The GC inlet temperature was set to 250 °C, and a split injection mode with a split ratio of 100:1 was applied. The oven temperature was maintained isothermally at 35 °C for 5 min, which was sufficient to achieve complete elution and baseline separation of all target compounds. Headspace conditions were set to 70 °C for the oven, 80 °C for the loop, and 90 °C for the transfer line. Samples were equilibrated in the headspace oven for 10 min prior to injection, and the injection duration was 0.5 min. The mass spectrometer was operated in electron impact (EI) ionization mode at an ionization energy of 70 eV. The source and quadruple temperatures were set to 230 °C and 150 °C, respectively. Initial method development and peak identification were carried out in full-scan mode (m/z 30–80) to verify peak identity and chromatographic separation. Subsequently, selected ion monitoring (SIM) mode was employed to improve sensitivity and quantitative performance. The monitored ions and corresponding retention time windows for each analyte are summarized in Table 1:

Table 1: Retention time windows and quantifier m/z values used for SIM mode for each component.

Component	Time Segments	m/z
Acetaldehyde	-1.20	29
Ethanol	1.20-1.48	31
1-propanol	1.48-1.80	31
Ethyl acetate	1.80-	43

Sample preparation & data acquisition

Standard mixtures were prepared to generate calibration and simulated reaction datasets. All solutions were transferred into 20 mL headspace vials and salted with 3.20 g of sodium chloride per vial to enhance analyte partitioning into the headspace. The vials were immediately sealed with Teflon/silicone crimp caps prior to analysis. Standard mixtures containing 1-propanol as an internal standard (IS) were prepared, with individual analyte concentrations varied between 50 and 1000 mg L^{-1} across four concentration levels. A full factorial experimental design was employed to generate all possible combinations of species. For each combination, four calibration standards were prepared. The concentration range was divided into four equal intervals, and Latin hypercube sampling was used to assign one concentration value from each interval to every species present in a given mixture. The sampled concentration values were rounded for accurate and practical pipetting. Solutions

containing only acetic acid were excluded from the HS–GC–MS dataset. Due to its significantly lower volatility, acetic acid was quantified separately using ion chromatography (IC), as acetic acid is the only ionic species among the target analytes. Nevertheless, acetic acid was included in all standard mixtures, as its presence can influence vapor–liquid equilibrium and thereby affect the headspace partitioning behavior of the volatile components. For mixtures containing all analytes, simulated electrooxidation samples were generated by assuming a representative bioethanol feed concentration of approximately 2 M. Random selectivities toward products were applied, which yielded 12 simulated electrooxidation samples and a total of 64 calibration samples. The peak area of each analyte was extracted and normalized to the peak area of the internal standard. These normalized peak areas were used as input variables for model training, while the corresponding known concentrations served as output variables.

Model Development

To analyze the effects of increasing model complexity and prior-induced model architectures on prediction accuracy and model training, 5 different models were trained: MLR, polynomial regression, ANN, prior-induced polynomial regression, and prior-induced branched ANN. The MLR model was used as the baseline for comparison with the more complex models. Polynomial regression and ANN models were also included to enable direct comparison with the prior-induced models. All models were evaluated using 5-fold cross-validation, and MAE and RMSE were used as metrics to compare predictive performance. Model complexities were additionally assessed using the AIC an BIC.

Baseline calibration models

MLR model was employed as a baseline model due to its simplicity, interpretability, and widespread use in chemometrics. The MLR model is expressed as:

$$\hat{y} = \mathbf{X}\beta \tag{1}$$

Where $\hat{y}$ denotes the predicted output vector, $\mathbf{X}$ is the input matrix, β is the regression coefficients. Polynomial regression and ANN models were also developed to consider potential nonlinear relationships between normalized areas and analyte concentrations, and for comparison with the proposed prior-induced models. Polynomial model extends the linear model by applying a polynomial feature expansion $\mathbf{\Phi}(.)$ of a degree 2:

$$\hat{y} = \mathbf{\Phi}(\mathbf{X})^{\mathrm{T}}\beta \tag{2}$$

Despite the nonlinear transformation of the inputs, the model is still linear with respect to the parameters, β. ANNs were used as a flexible nonlinear models capable of capturing complex interactions among variables. The general expression for a feedforward ANN with L layers is given by:

$$\hat{y} = \boldsymbol{W}_L\sigma_{L-1}\left(\boldsymbol{W}_{L-1}\sigma_{L-2}\left(\ldots\sigma_1(\boldsymbol{W}_1\mathbf{X})\right)\right) \tag{3}$$

Where $\boldsymbol{W}_L$ represents the weight matrices (including bias terms) of layer L, σ_L denotes the activation function of the corresponding layer.

Linear and polynomial regression models were implemented using the scikit-learn library (version 1.7.2), while prior-induced polynomial regression and ANN models were developed using PyTorch (version 2.5.1) in Python. All models were trained and evaluated using 5-fold cross-validation to ensure robust performance assessment. For all models, normalized peak areas were used as input variables, yielding a total of 3 input features. Each model predicts 3 outputs, corresponding to the concentrations of the respective species. All input and output variables were normalized in the range 0-1. ANN training employed the mean squared error (MSE) as the loss function. Hyperparameters of the ANN models were selected using a manual exploration strategy to identify an optimized network architecture and training configuration considering dataset size. For this, a nested cross-validation strategy was adopted, where hyperparameter optimization was conducted using an additional 5-fold cross-validation within the training set of each k-fold. The activation functions evaluated included the rectified linear unit (ReLU) and the hyperbolic tangent (tanh). The number of hidden layers was varied between 1 and 3, and the number of neurons per layer ranged from 3 to 36. Model training was performed using the Adam optimizer, with learning rates manually explored in the range of 10^{-4} and 10^{-2}. Xavier weight initialization were applied to enhance training stability and convergence.

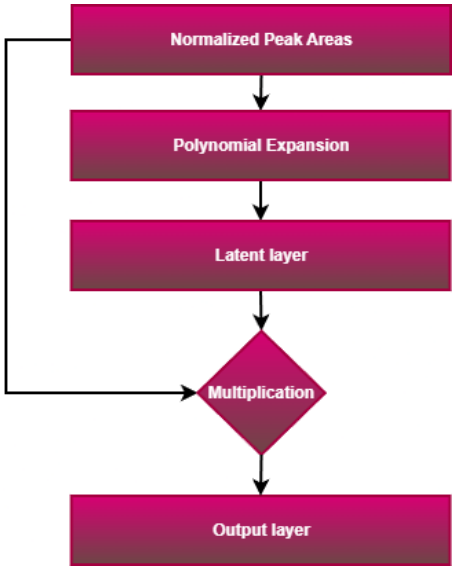

Figure 1. Physical prior-induced polynomial regression workflow.

Integration of physical prior

The prior-induced polynomial regression model was implemented in PyTorch using 3 input layers in the input layer. Polynomial feature expansion was applied to the inputs in the first layer. This layer was followed by the formation of intermediate terms. Then, these intermediate terms were multiplied with the input layer and linearly mapped to the output layer, yielding the final outputs. The resulting architecture is visualized in the Figure 1. Model training was performed using the MSE loss function and the Adam optimizer, with a fixed learning rate of 10^{-3}.

The ANN model was constructed following a similar prior-integration strategy, in which the original input variables were multiplied with the intermediate terms. To make ANN specialized for each of the output, the ANN architecture was partitioned into 3 separate sub-branches, each corresponding to an individual output. The model incorporates a shared base layer to capture common features across all outputs, followed by dedicated branch layers, which aims output-wise refinement and increased predictive performance. This branched ANN architecture, which balanced the shared information and output-specific modelling with an incorporation of physical prior is illustrated in the Figure 2.

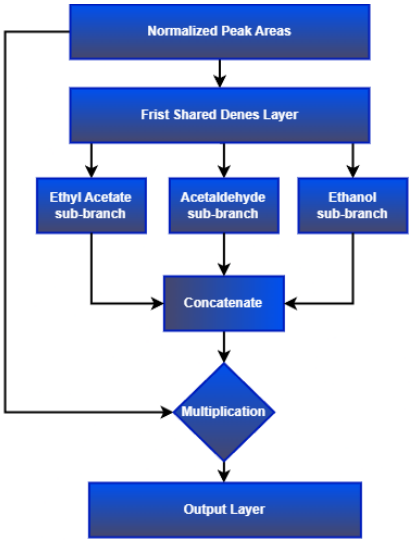

Figure 2. Physical prior-induced sub-branched ANN model.

Model Evaluation Criteria

AIC and BIC are widely used model selection criteria in machine learning that balances goodness of fit against model complexity. The two criteria differ in the way model complexity is penalized. In this study, both metrics were used to evaluate the proposed models in order to avoid selection of unnecessarily complex calibration models. For regression models, AIC is [11]:

$$AIC = n \bullet \ln\left(\frac{RSS}{n}\right) + 2k \quad (1)$$

Where n is the total number of data points used for fitting, and k is the number of parameters, which penalizes the model complexity. The first term in the equation indicates goodness of fit, and RSS is the residual sum of squares, and calculated across all components since there are 3 outputs:

$$RSS = \sum_{i=1}^{n} \sum_{j=1}^{3} \left(y_{ij} - \hat{y}_{ij}\right)^2 \quad (2)$$

In this case, AIC is calculated fold-wise and average of the 5-fold is reported to compare the models.

BIC is derived from Bayesian probability theory, and used for model selection [12]. It balances likelihood with a sample-size penalty. Like AIC, the first term evaluates the goodness of fit. BIC penalizes the number of parameters more strongly than AIC with the multiplication of $\ln(n)$ in the second term, which promotes simpler models.

$$BIC = n \bullet \ln\left(\frac{RSS}{n}\right) + k \bullet \ln(n) \quad (3)$$

For ANNs, BIC strictly limits the number of variables and favors linear models like polynomial regression with a low degree.

RESULTS & DISCUSSION

Figure 3 shows the relationship between concentration and the response behaviour of each analyte, where the primary y-axis is normalized area and the secondary y-axis corresponds to the raw peak area. The application of an internal standard has a very clear effect mitigating fluctuations between injections and instrumental variability, and improves stability and robustness significantly. However, Figure 3 also reveals that application of internal standard is not sufficient to explain relation with a perfectly linear trend, as there is still noticeable response variability.

Baseline Model Performance

Averaged over 5-fold cross-validation, MLR model yields a component-wise MAE and RMSE values of 29.26 and 35.40 for ethanol, 45.16 and 55.55 for acetaldehyde, and 49.75 and 58.00 for ethyl acetate. These results serve as a baseline for comparison with more flexible calibration models.

Parity plots for the test results of each k-fold obtained using the MLR model are shown in Figure 4 for each component separately. A clear model bias is observed in all plots, where predictions are above the parity line in the center of the range and below the parity line at the extremes. In addition, the MLR model produces non-zero predictions at zero concentration, this phenomenon is particularly evident for ethyl acetate predictions.

Standard polynomial regression with a second degree expansion results in a significant improvement in both metrics. The component-wise MAE and RMSE values are given in Table 2. The significant improvement in prediction accuracy indicates that the calibration model benefits from incorporating nonlinear relationships between the response and analyte concentrations.

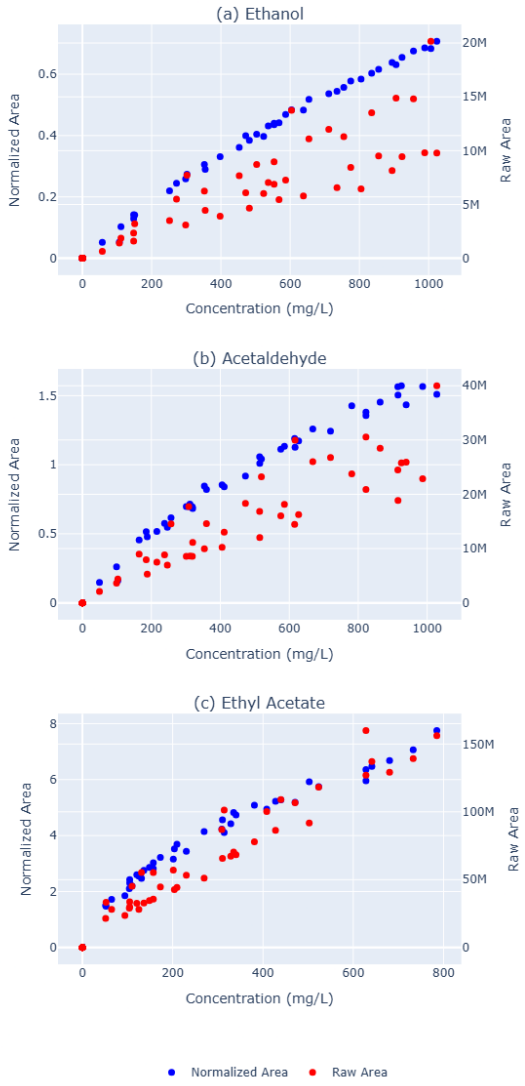

Figure 3. Standard concentrations and raw and internal standard normalized peak areas for (a) ethanol, (b) acetaldehyde, and (c) ethyl acetate.

The ANN model also exhibits predictive performance similar to polynomial regression. As shown in Table 3, the ANN yields a slightly higher MAE and RMSE for acetaldehyde, while achieving a nearly 10% decrease in MAE and RMSE for ethyl acetate. Overall, both polynomial regression and ANN outperform the MLR model, which indicates the importance of incorporating nonlinear relationships between the input variables and analyte concentrations.

Table 2: Component-wise MAE and RMSE for test set predictions, averaged over 5-fold cross-validation, obtained using polynomial regression.

Component	MAE	RMSE
Ethanol	10.11	15.03
Acetaldehyde	18.30	29.92
Ethyl acetate	15.17	23.89

Table 3: Component-wise MAE and RMSE for test set predictions, averaged over 5-fold cross-validation, obtained using ANN.

Component	MAE	RMSE
Ethanol	10.30	14.50
Acetaldehyde	19.08	30.46
Ethyl acetate	13.50	20.81

Table 4: Component-wise MAE and RMSE for test set predictions, averaged over 5-fold cross-validation, obtained using prior-induced polynomial regression.

Component	MAE	RMSE
Ethanol	7.50	12.55
Acetaldehyde	18.56	33.18
Ethyl acetate	12.04	20.50

Table 5: Component-wise MAE and RMSE for test set predictions, averaged over 5-fold cross-validation, obtained using polynomial regression prior-induced branched ANN model.

Component	MAE	RMSE
Ethanol	7.98	13.45
Acetaldehyde	17.37	29.85
Ethyl acetate	12.84	23.50

Table 6: Number of parameters, k, AIC, and BIC values for each model.

Model	k	AIC	BIC
MLR	12	332	352
Polynomial regression	30	302	352
Fully connected ANN	115	475	545
Prior-induced polynomial regression	30	316	370
Prior-induced branched ANN	155	665	803

Prior-Induced Models

The prior-induced polynomial regression and branched ANN models were evaluated using the same validation procedure as the baseline models. Incorporating prior information into the polynomial regression framework results in improved prediction accuracy for ethanol and ethyl acetate, while it yields comparable performance for acetaldehyde across both error metrics.

The component-wise MAE and RMSE values obtained using this model are summarized in Table 4.

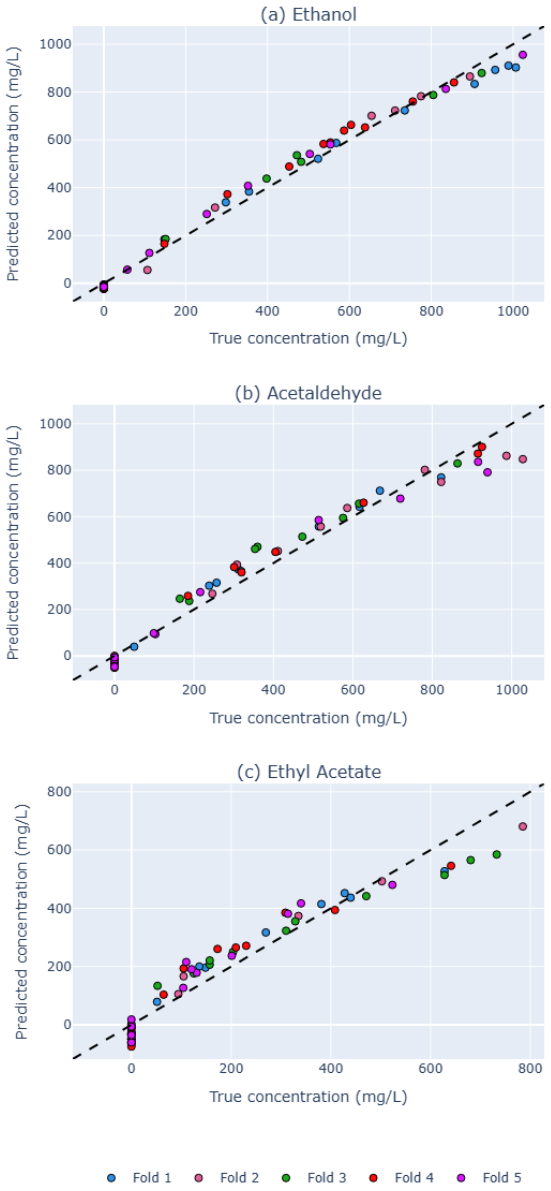

Figure 4. Parity plots for the test set of each k-fold using MLR for (a) ethanol, (b) acetaldehyde, and (c) ethyl acetate.

The prior-induced branched ANN model demonstrates consistent improvement across all metrics compared the fully connected ANN model, however, when it is compared with prior-induced polynomial regression model, branched ANN achieves slightly better prediction accuracy for acetaldehyde, while having marginally higher errors for ethanol and ethyl acetate. These results suggest that embedding prior knowledge into the network enhances predictive performance, while still having an efficient training process.

Model Complexity

Table 6 compares AIC and BIC values obtained for MLR, polynomial regression, fully connected ANN, and their prior-induced counterparts. Both AIC and BIC evaluate model performance by balancing goodness of fit against model complexity through an explicit penalty on the number of parameters. Among all evaluated models standard polynomial regression exhibits the lowest AIC and one of the lowest BIC values, which indicates that the improvement in predictive performance justifies the additional complexity relative to the MLR baseline. The fully connected ANN and prior-induced ANN results in substantially higher AIC and BIC, showing that large number of trainable parameters in ANNs are strongly penalized. The prior-induced polynomial regression has slightly increased AIC and BIC compared to standard polynomial regression. Since both models contain the same number of trainable parameters, this increase can be attributed to a higher RSS for the prior-induced formulation.

When all models are considered, ANN-based models do not yield proportional improvements in prediction accuracy despite their increased complexity, while the prior-induced polynomial regression model achieves predictive performance close to the ANN-based models, while having AIC and BIC comparable to MLR model, which makes it a good alternative in this specific use-case. However, both prior-induced polynomial regression model and branched-ANN model can provide efficient frameworks for incorporating physical prior information into calibration models for similar or more complex systems affected by matrix effects.

CONCLUSION

In conclusion, data-driven calibration models were developed for the accurate quantitative analysis of ethanol electrooxidation products using HS–GC–MS calibration data. MLR, polynomial regression, and fully-connected ANN were employed as baseline calibration models, while prior-induced polynomial regression and ANN models were developed by incorporating a simple physical constraint: The principle that a zero response corresponds to zero analyte concentration. Both polynomial regression and ANN models substantially outperformed the MLR baseline, achieving approximately 60–70% reductions in component-wise MAE and RMSE. Polynomial regression exhibited slightly better performance in acetaldehyde prediction, whereas ANN yielded marginally lower MAE and RMSE for ethyl acetate. Incorporation of prior information with polynomial regression further reduced MAE and RMSE by approximately 20–25% for ethanol and acetaldehyde. The prior-induced branched ANN achieved predictive accuracy comparable to that of the prior-induced polynomial regression model. However, both fully connected and branched ANN models

exhibited AIC and BIC values approximately twice those of the polynomial models, indicating that the ANN architectures are overly complex for the present problem. Overall, the proposed methodology demonstrates strong potential for application to more complex solution systems by effectively integrating prior physical knowledge and nonlinear relationships into model training.

ACKNOWLEDGEMENTS

Prof. Leblebici gratefully acknowledges the support of KU Leuven C2 project Scalable performant solar cell module for CO2 Reduction Coupled with Organic Synthesis (SPROUT) C2E/23/010

REFERENCES

1. St?ber M, Reemtsma T. Evaluation of three calibration methods to compensate matrix effects in environmental analysis with LC-ESI-MS. Analytical and Bioanalytical Chemistry 378:910-916 (2004). https://doi.org/10.1007/s00216-003-2442-8
2. Cheng WL, Markus C, Lim CY, Tan RZ, Sethi SK, Loh TP, . Calibration practices in clinical mass spectrometry: review and recommendations. Ann Lab Med 43:5-18 (2022). https://doi.org/10.3343/alm.2023.43.1.5
3. Moosavi SM, Ghassabian S. Linearity of calibration curves for analytical methods: a review of criteria for assessment of method reliability. Calibration and Validation of Analytical Methods - A Sampling of Current Approaches : (2018). https://doi.org/10.5772/intechopen.72932
4. Olivieri AC. Chemometrics and multivariate calibration. Introduction to Multivariate Calibration :1-26 (2024). https://doi.org/10.1007/978-3-031-64144-2_1
5. Veiga-del-Baño JM, Oliva J, Cámara MÁ, Andreo-Martínez P, Motas M. Matrix-matched calibration for the quantitative analysis of pesticides in pepper and wheat flour: selection of the best calibration model. Agriculture 14:1014 (2024). https://doi.org/10.3390/agriculture14071014
6. Chen HY, Chen C. Evaluation of calibration equations by using regression analysis: an example of chemical analysis. Sensors 22:447 (2022). https://doi.org/10.3390/s22020447
7. Ferreira V, Herrero P, Zapata J, Escudero A. Coping with matrix effects in headspace solid phase microextraction gas chromatography using multivariate calibration strategies. Journal of Chromatography A 1407:30-41 (2015). https://doi.org/10.1016/j.chroma.2015.06.058
8. Campmajó G, Saez-Vigo R, Saurina J, Núñez O. High-performance liquid chromatography with fluorescence detection fingerprinting combined with chemometrics for nut classification and the detection and quantitation of almond-based product adulterations. Food Control 114:107265 (2020). https://doi.org/10.1016/j.foodcont.2020.107265
9. Brusamarello CZ, , Di Domenico M, Da Silva C, de Castilhos F. A COMPARATIVE STUDY BETWEEN MULTIVARIATE CALIBRATION AND ARTIFICIAL NEURAL NETWORK IN QUANTIFICATION OF SOYBEAN BIODIESEL. Rev.Mex.Ing.Quim. 19:123-132 (2019). https://doi.org/10.24275/rmiq/bio579
10. Gholivand MB, Jalalvand AR, Goicoechea HC, Gargallo R, Skov T, Paimard G. Combination of electrochemistry with chemometrics to introduce an efficient analytical method for simultaneous quantification of five opium alkaloids in complex matrices. Talanta 131:26-37 (2015). https://doi.org/10.1016/j.talanta.2014.07.053
11. Cavanaugh JE, Neath AA. The akaike information criterion: background, derivation, properties, application, interpretation, and refinements. WIREs Computational Stats 11: (2019). https://doi.org/10.1002/wics.1460
12. Neath AA, Cavanaugh JE. The bayesian information criterion: background, derivation, and applications. WIREs Computational Stats 4:199-203 (2011). https://doi.org/10.1002/wics.199