Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i2.417

Research Article

Interpreting Reactive Species Interdependencies in Plasma-Activated Saline: An Exploratory Multivariate Workflow

Nilton Francelosi Azevedo Neto[1]; Orisson Ponce Gomes[2]; Lucas Pereira Piedade[3]; Felipe Souza Miranda[4]; Rodrigo Savio Pessoa[5]

[1] Aeronautics Institute of Technology, Brazil, São José dos Campos, nilton.azevedo@unesp.br
[2] Universite de Pau et des Pays de L'Adour, France, Pau, orissongomes@gmail.com
[3] École de Technologie Supérieure, Canada, Montreal, lucas.piedade@gmail.com
[4] Aeronautics Institute of Technology, Brazil, São José dos Campos, miranndafs@gmail.com
[5]Aeronautics Institute of Technology, Brazil, São José dos Campos, rodrigospessoa@gmail.com

Correspondence should be addressed to Nilton Francelosi Azevedo Neto; nilton.azevedo@unesp.br

Received 10 March 2025; Accepted 15 June 2025; Published 31 July 2025

Abstract:
Introduction: Interpreting Plasma-Activated Saline (PAS) is challenging because reactive oxygen, nitrogen, and chlorine species are strongly interdependent, making it difficult to isolate their individual contributions to physicochemical properties such as Oxidation-Reduction Potential (ORP). This study explores these relationships using a multivariate chemometric workflow. **Method**: A small experimental dataset comprising 10 PAS observations generated by a serial DBD-GAPJ reactor was analyzed using Exploratory Data Analysis and Multiple Linear Regression. ORP was modeled as a function of O3, H2O2, HClO, NO3−, and NO2−. Model interpretation was supported by Variance Inflation Factor analysis, standardized coefficients, Leave-One-Out Cross-Validation, and residual diagnostics. **Results and Discussion**: H2O2 showed the strongest bivariate correlation with ORP (r = 0.86). The regression model achieved $R^2 = 0.84$ and adjusted $R^2 = 0.65$, but LOOCV produced $Q^2 = -0.33$, indicating poor out-of-sample prediction. Strong multicollinearity among reactive species complicated coefficient interpretation, and none of the individual predictors reached conventional statistical significance. Standardized coefficients identified H2O2 as the strongest relative contributor, while O3 and HClO remained chemically plausible interdependent contributors. **Conclusion**: The proposed workflow is valuable for interpreting small, collinear chemical datasets, but the results should be regarded as hypothesis-generating rather than predictive, emphasizing the importance of standardization, cross-validation, and multicollinearity diagnostics.

Keywords: Plasma-Activated Saline; Reactive Oxygen and Nitrogen Species; Multiple Linear Regression; Multicollinearity; Oxidation-Reduction Potential.

1. Introduction

Plasma-activated liquids, such as Plasma-Activated Saline (PAS), are emerging as a versatile, low-cost, and environmentally friendly technology with significant potential across fields ranging from water disinfection to biomedicine [1], [2], [3], [4], [5]. The remarkable efficacy of PAS in applications such as microbial inactivation and cancer therapy is largely attributed to its potent oxidative properties, which are commonly quantified by a key physicochemical parameter: the Oxidation-Reduction Potential (ORP). A high ORP directly reflects a solution's ability to disrupt biological systems, making it a critical metric for evaluating and optimizing the therapeutic or sanitizing potential of PAS [6]. This elevated potential arises from a complex mixture of long-lived reactive oxygen and nitrogen species (RONS), including nitrate (NO_3^-), nitrite (NO_2^-), hydrogen peroxide (H_2O_2), and ozone (O_3), as well as reactive chlorine species (RCS) like HClO, which are formed in saline solutions [7], [8], [9].

The specific chemical composition of PAS is highly dependent on the plasma source technology and the gases used in the discharge, as different reactor types produce varying concentrations of reactive species [10]. Among these, the serial association of a Dielectric Barrier Discharge (DBD) with a Gliding Arc Plasma Jet (GAPJ) has shown

particularly promising results for activating saline water, and was therefore selected for the detailed analysis presented in this study [11], [12].

However, clarifying the quantitative relationship between the solution's chemical composition and its resulting oxidative potential remains a significant scientific challenge. A central issue is identifying which species serve as the primary drivers of this crucial potential. The answer is not straightforward: overall ORP depends not only on the concentration of each species but also on its intrinsic oxidizing strength (E°) [7]. As a result, simple bivariate analysis can be misleading, making it difficult to isolate the individual contribution of each component in a multivariate system. To address this challenge, the present study employs a chemometric workflow that combines exploratory data analysis with multivariate modeling to deconstruct these complex chemical interdependencies [13], [14], [15], [16].

This work, therefore, presents a systematic, exploratory analytical workflow to analyze data from PAS generated by a serial DBD-GAPJ reactor. Given the availability of only ten activation-time observations, the analysis is conceived as a hypothesis-generating case study rather than as a fully validated predictive framework, and its multivariate tools, including correlation analysis and multiple linear regression, are combined with diagnostic procedures to assess the reliability of the resulting model. The primary objective is to investigate the relative influence of RONS and RCS on the solution's ORP, moving beyond simple correlations to distinguish plausible chemical drivers from statistical artifacts such as multicollinearity. In doing so, this study aims to provide clear, actionable insights that can inform the future design and optimization of plasma-activated liquids, while illustrating a workflow for interpreting complex chemical datasets when multicollinearity and small sample size limit definitive causal conclusions.

2. Method

A. Exploratory Data Analysis and Regression Modeling

The dataset utilized in this study, which characterizes the reactive species concentrations and physicochemical properties of Plasma-Activated Saline (PAS), was sourced from previously reported experimental work [11]. The present analysis focuses on the development of a chemometric framework designed to decouple the complex chemical interdependencies inherent in this system. Unlike traditional bivariate assessments, this multivariate approach aims to distinguish between the intrinsic chemical drivers of ORP and statistical artifacts arising from collinear generation pathways of RONS. Given the limited size of the dataset (n = 10 observations), the analysis is presented as an exploratory, hypothesis-generating investigation rather than as a validated predictive model [17], and dedicated diagnostic procedures were implemented to assess the reliability of the regression results.

The analysis was conducted using a computational routine developed in Python (v3.13.6), leveraging its open-source scientific libraries: Pandas for data manipulation, Matplotlib and Seaborn for visualization, and Scikit-learn for statistical modeling [18], [19], [20]. To ensure a reproducible and structured approach, the entire process was implemented within an exploratory analytical workflow (Figure 1), which guides the study sequentially from data preparation and exploratory data analysis (EDA) through multivariate modeling, statistical diagnostics, and final knowledge synthesis.

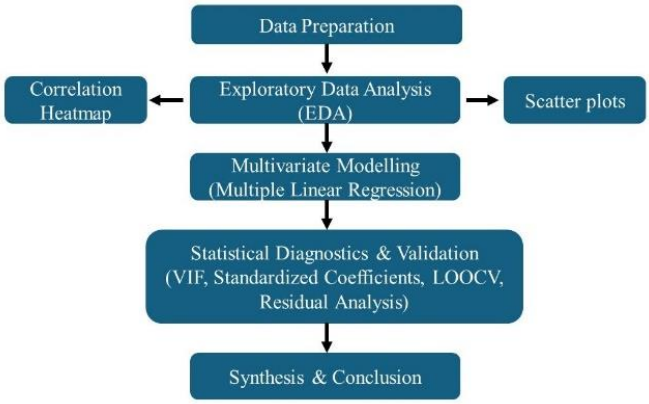

Figure 1. The exploratory analytical workflow employed in this study. The framework guides the analysis sequentially from data preparation and exploratory analysis (EDA) through multivariate modeling, statistical diagnostics, and final knowledge synthesis.

This framework guides the analysis of the dataset used in this study (summarized in Table 1), obtained from a previous publication [11], from initial data exploration to final statistical diagnostics.

Table 1. Physicochemical properties and reactive species concentrations of Plasma-Activated Saline (PAS) at different activation times. This dataset, previously published in [11], served as the basis for the exploratory data analysis, regression modeling, and statistical diagnostics presented in this work.

Time (min)	pH	k (mS/cm)	ORP (mV)	H_2O_2 (mg/L)	O_3 (mg/L)	HClO (mg/L)	NO_3^- (mg/L)	NO_2^- (mg/L)
1	3.67	13.61	180.6	9.5	0.02	0.02	25	10
2	3.38	39	209	22	0.05	0.02	25	10
4	3.16	37.14	224.4	27	0.06	0.09	50	20
6	3.07	37.68	237	31	0.1	0.7	100	40
8	2.79	38.11	239.2	29	0.47	1.3	100	40
10	2.72	41.43	229.6	30	0.52	1.51	250	40
15	2.71	38.21	235	36	0.82	3.88	250	40
20	2.58	39.31	243.1	37	1.58	3.52	250	40
25	2.57	41.63	261	36	2	2.87	250	20
30	2.5	41.43	222.1	37	2	2.87	100	20

B. Exploratory Data Analysis

The initial stage of the analysis focused on characterizing the overall structure of the data and identifying key relationships among the measured variables. A Pearson correlation matrix was first computed to quantify the linear relationship between all variable pairs (e.g., ORP vs. O_3, pH vs. HClO), and the results were visualized as a heatmap to enable rapid identification of strong correlations. To further examine these relationships, the most significant correlations were subsequently visualized through bivariate scatter plots with fitted linear trendlines and 95% confidence intervals.

C. Multiple Linear Regression Modeling

Following the exploratory analysis, a multiple linear regression model was developed with the primary objective of interpretation. The aim was to build a model that could quantify the relative influence of each predictor variable on the ORP. This relationship is described by the linear equation, where the predicted ORP is a function of the concentrations ([X]) of the five measured reactive species:

$$ORP=\beta_0+\beta_1[O_3]+\beta_2[H_2O_2]+\beta_3[HClO]+\beta_4[NO_3^-]+\beta_5[NO_2^-]+\varepsilon \quad (1)$$

In this model, β_0 represents the intercept, while the regression coefficients β_1, β_2, β_3, β_4, and β_5 quantify the individual influence of each corresponding species on the ORP, and ε represents the model's residual error. The model was fitted using ordinary least squares (OLS) regression, implemented independently with Scikit-learn and Statsmodels to obtain both the fitted coefficients and their associated inferential statistics (standard errors, 95% confidence intervals, and p-values). Given the small sample size (n = 10) relative to the number of predictors, the calculated coefficients were treated as exploratory indicators of relative association rather than as confirmed, individually significant effects, pending the diagnostic procedures described in Section D.

D. Statistical Diagnostics and Model Validation

To assess the reliability and interpretability of the regression model, several diagnostic procedures were implemented. First, multicollinearity among predictors was quantified using the Variance Inflation Factor (VIF) for each species, with VIF values above 5 and 10 conventionally taken to indicate moderate and severe multicollinearity, respectively [21]. Second, because the five predictors span markedly different concentration ranges, standardized regression coefficients (β) were computed by z-score standardizing all predictors and the response variable prior to fitting, allowing a scale-independent comparison of relative influence. Third, the predictive performance of the model was evaluated using Leave-One-Out Cross-Validation (LOOCV), in which each observation was iteratively excluded from the training set and predicted from a model fitted on the remaining nine observations; predictive performance was summarized using the cross-validated coefficient of determination (Q^2) and root-mean-square error (RMSE), and compared against the corresponding in-sample metrics [22]. Given the small sample size relative to the number of predictors, these diagnostics follow established recommendations for regression analysis under limited subjects-per-variable ratios [17]. Finally, residual diagnostics were performed to evaluate the assumptions underlying the regression model, including the Shapiro-Wilk test for normality of residuals, the Durbin-Watson statistic for autocorrelation, the correlation between residuals and activation time to assess potential temporal confounding, and Cook's distance to identify influential observations. All statistical analyses were implemented in Python (v3.13.6) using Pandas, Scikit-learn, Statsmodels, and SciPy.

3. Result and Discussion

To investigate the interdependencies among the physicochemical properties and reactive species, a Pearson correlation matrix was calculated (**Figure 2**). The analysis highlighted several chemically meaningful correlations, most notably a strong negative correlation between pH and activation time (r = - 0.86), confirming the progressive acidification of the solution. This trend is consistent with the strong positive correlations observed between activation time and the concentrations of key oxidizing species, namely O_3 (r = + 0.98) and HClO (r = + 0.84), indicating their accumulation throughout the process.

The analysis of the ORP, however, revealed a more complex dynamic. While all oxidizing species showed a positive correlation with ORP, the strongest linear relationship was unexpectedly observed with H_2O_2 (r = + 0.86). The correlations with HClO (r = + 0.61) and O_3 (r = + 0.56) were moderate in comparison. This counterintuitive outcome from the exploratory analysis highlights the challenge of interpreting bivariate correlations in a multivariate system and sets the stage for the necessity of a more robust model to disentangle these convoluted effects. Additionally, a strong negative correlation was found between pH and HClO (r = -0.85), further corroborating that the generation of this reactive chlorine species is a key driver of the medium's acidification [6], [23], [24].

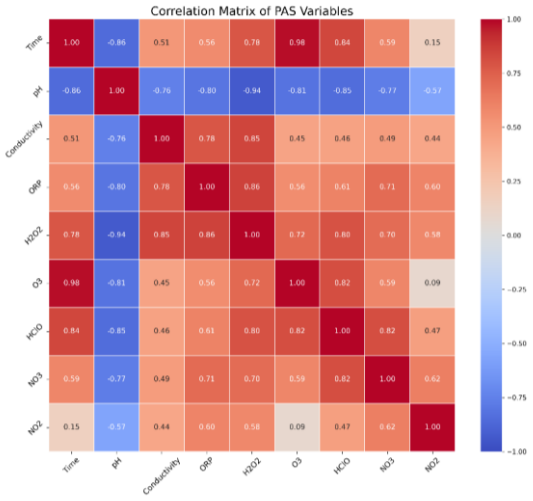

Figure 2. Pearson correlation matrix of the physicochemical properties and reactive species concentrations in Plasma-Activated Saline (PAS). The values in each cell represent the Pearson correlation coefficient (r) between the corresponding variables. The color scale on the right indicates the strength and direction of the correlation, with strong positive correlations approaching +1 (red) and strong negative correlations approaching -1 (blue).

The correlation matrix displayed as a heatmap in **Figure 2**, revealed several strong linear relationships. To further investigate and visually validate these findings, scatter plots were generated for the most significant variable pairs (**Figure 3**).

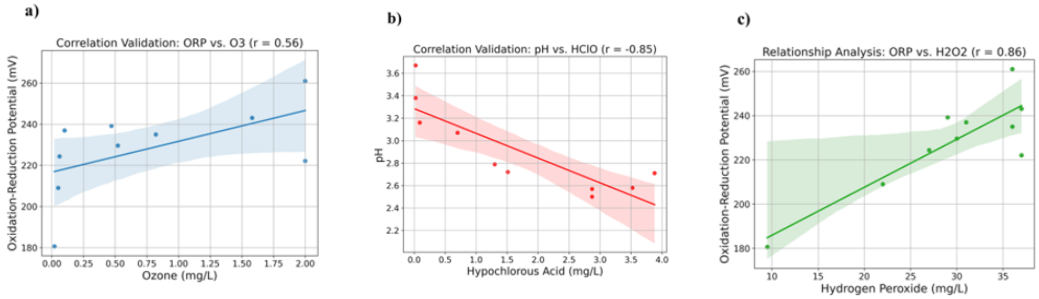

Figure 3. Visual validation of key linear correlations identified in the correlation matrix (**Figure 2**). Each subplot displays the data points, the linear regression trendline (solid line), and the 95% confidence interval (shaded area). To quantify the strength of the linear relationship, the Pearson correlation coefficient (r), corresponding to the value in the heatmap, is indicatedin each subplot title. a) Moderate positive correlation between ORP and Ozone (O_3), b) strong negative correlation between pH and Hypochlorous Acid (HClO), and c) strong positive correlation between ORP and Hydrogen Peroxide (H_2O_2).

Figure 3a depicts the direct relationship between ozone concentration and the solution's oxidative state. A moderate, positive linear correlation is observed between ORP and O_3 (r = +0.56). The data points show some scatter around the linear regression trendline, with a relatively wide 95% confidence interval, suggesting that while ozone contributes to the oxidative capacity, its direct bivariate relationship with ORP is not as strong as might be initially expected.

Figure 3b investigates the concurrent process of acidification, showing a strong negative correlation between the solution's pH and the HClO concentration (r = - 0.85). The close alignment of the data with the regression line and the narrow confidence interval underscore the robustness of this relationship. These results provide strongevidence that HClO formation is a major contributor to the mechanisms responsible for acidification.

Figure 3c is presented for comparative purposes, examining the contribution of H_2O_2 to ORP. A strong positive correlation is evident (r = + 0.86), with data points tightly clusteredaround the regression line and a narrow confidence interval. This indicates a robust direct linear relationship between H_2O_2 concentration and ORP, suggesting H_2O_2 is a primary individual contributor to the solution's oxidative capacity in this bivariate context. This observation, when compared with the moderate correlation of O_3 (Figure 3a), highlights the complex interdependencies among the generated oxidizing species, necessitating a multivariate approach for a comprehensive understanding [25].

Although the exploratory analysis revealed strong linear correlations, it cannot resolve the individual contribution of each species in a multivariate system. Therefore, to quantify the relative influence of each reactive species on ORP, a multiple linear regression model was developed. Beyond fitting the model to the available data, its performance was evaluated both in-sample and through Leave-One-Out Cross-Validation (LOOCV), given the limited size of the dataset (n = 10). The full set of regression coefficients, standard errors, confidence intervals, p-values, and variance inflation factors (VIF) is reported in **Table 2**. The model's performance and the interpretation of its coefficients are shown in **Figures 4**, **5**, and **6**.

Table 2. Multiple linear regression results for the prediction of ORP as a function of O_3, H_2O_2, HClO, NO_3^-, and NO_2^- concentrations (n = 10). The table reports unstandardized coefficients (B), standardized coefficients (β), standard errors, 95% confidence intervals, p-values, and variance inflation factors (VIF) for each predictor. Standardized coefficients account for the markedly different concentration ranges of the predictors and enable direct comparison of their relative influence on ORP. None of the predictors reached conventional statistical significance (p > 0.10), consistent with the severe multicollinearity indicated by the VIF values and the limited residual degrees of freedom associated with the small sample size.

Reactive Species	Unstandardized Coefficient (B)	Standardized Coefficient (β)	Std. Error	95% CI (Lower)	95% CI (Upper)	p-value	VIF
H_2O_2	2.213	0.875	1.0888	-0.8103	5.2357	0.1119	4.7652
NO_3^-	0.105	0.480	7.7198	-31.9924	10.8748	0.2432	7.3182
O_3	6.227	0.232	0.0859	-0.1331	0.3438	0.2873	3.9288
NO_2^-	0.188	0.114	13.6754	-31.7425	44.1958	0.6725	6.6637
HClO	-10.559	-0.730	0.6335	-1.5708	1.947	0.7813	3.8003

As shown in **Figure 4a**, the model achieved an in-sample coefficient of determination of $R^2 = 0.84$ (adjusted $R^2 = 0.65$), indicating that the five predictors jointly account for a substantial portion of the variance in ORP within the observed data. However, when the model was evaluated under LOOCV, in which each observation was predicted from a model trained on the remaining nine data points, the predictive coefficient of determination dropped to $Q^2 = -0.33$ (Figure 4b), with the root-mean-square error increasing from 8.1 mV in-sample to 23.8 mV under cross-validation. A negative Q^2 indicates that the model performs worse than simply predicting the sample mean when applied to unseen data. This gap between in-sample fit and out-of-sample performance is consistent with the small sample size relative to the number of predictors (five predictors and an intercept fitted to ten observations, leaving only four residual degrees of freedom) and indicates that the regression model should be interpreted as an exploratory, hypothesis-generating tool rather than as a validated predictive model.

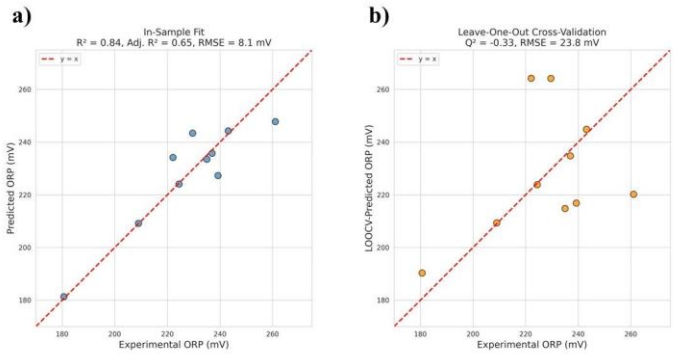

Figure 4. Evaluation of the multiple linear regression model's predictive performance, comparing in-sample fit with out-of-sample cross-validation. (a) Experimental versus model-predicted ORP values using the full dataset (n = 10). The dashed red line represents perfect agreement (y = x). The model achieved $R^2 = 0.84$ (adjusted $R^2 = 0.65$, RMSE = 8.1 mV), indicating a strong in-sample fit. (b) Experimental versus predicted ORP values obtained through Leave-One-Out Cross-Validation (LOOCV), in which each observation was predicted using a model trained on the remaining nine data points. The negative predictive coefficient of determination ($Q^2 = -0.33$, RMSE = 23.8 mV) indicates that the model has no genuine predictive capability beyond the training sample, underscoring the exploratory nature of the analysis given the small dataset size.

As shown in **Figure 5a**, the multiple linear regression model assigns a large positive raw coefficient to O3 and a large negative raw coefficient to HClO. The negative coefficient for HClO should be interpreted, at least in part, as a statistical consequence of the strong multicollinearity between these species, rather than reflecting a direct chemical effect on ORP. The strong positive correlation between O_3 and HClO (r = + 0.82, Figure 2; VIF = 6.7 and 7.3, respectively, **Table 2**) indicates their concurrent generation. This is consistent with established chemical pathways, where in the acidic environment produced by the plasma, O_3 is consumed during the oxidation of chloride ions, a key process leading to the formation of various RCS, including HClO [6], [26], [27]. Consequently, the model is unable to reliably separate the independent effects of O_3 and HClO on ORP, and compensates mathematically by assigning much of the positive influence to one species and a compensatory negative coefficient to the other.

This mathematical compensation, however, raises an important caveat for the interpretation of coefficient magnitude: because O_3, H_2O_2, HClO, NO_3^-, and NO_2^- span very different concentration ranges (e.g., O_3: 0.02–2 mg/L; H_2O_2: 9.5–37 mg/L), raw regression coefficients are not directly comparable across predictors and cannot, by themselves, establish a hierarchy of chemical influence. To address this, standardized coefficients (β) were calculated by z-score standardizing all predictors and the response variable prior to fitting the model (Figure 5b). Under this standardized representation, H_2O_2 (β = + 0.88) emerges as the predictor with the largest relative influence on ORP, followed by HClO (β = - 0.73) and NO_3^- (β = + 0.48), while O_3 (β = + 0.23) shows a comparatively modest standardized effect despite its large raw coefficient. This indicates that the apparent dominance of O_3 in the raw-coefficient model is, at least in part, an artifact of its narrow concentration range rather than solely reflecting its individual influence on ORP. In datasets that remain small or strongly collinear, regularized regression approaches such as ridge regression [28] or partial least squares regression [29] offer promising alternatives to ordinary least squares, as they are specifically designed to produce stable coefficient estimates under multicollinearity, and could be explored in future extensions of this analysis.

Importantly, none of the five predictors reached conventional statistical significance at the 95% confidence level (all $p > 0.10$; Table 2), with H_2O_2 showing the smallest, though still non-significant, p-value ($p = 0.112$). This lack of individual significance is consistent with the severe multicollinearity among predictors (VIF ranging from 3.80 to 7.32) and the limited residual degrees of freedom (four, given five predictors and an intercept fitted to ten observations). Consequently, while the overall model explains a substantial share of the variance in ORP ($R^2 = 0.84$), the individual regression coefficients, whether raw or standardized, should not be interpreted as establishing statistically confirmed, independent effects of any single species on ORP. They are best regarded as descriptive

indicators of relative association within a collinear system, to be interpreted jointly with chemical domain knowledge rather than as isolated statistical evidence.

Considered together with the chemical principles governing the system's oxidative potential, this result should not be interpreted as ruling out a meaningful contribution from O_3. The established electrochemical hierarchy, governed by standard reduction potentials (E°), confirms O_3 (E° = +2.07 V) as a stronger oxidant than both HClO (E° = +1.49 V) and H_2O_2 (E° = +1.78 V) [30], and O_3 remains a chemically plausible contributor to ORP. However, given the strong multicollinearity between O_3 and HClO and the divergence between raw and standardized coefficients, the present dataset does not allow the individual contributions of these two species to be statistically disentangled with confidence. Rather than establishing a definitive chemical hierarchy, the regression analysis is best interpreted as providing quantitative, hypothesis-generating evidence that H_2O_2, O_3, and HClO are all statistically relevant, interdependent contributors to ORP, with their relative individual weights remaining an open question warranting confirmation through dedicated experimental designs (e.g., factorial variation of individual species concentrations).

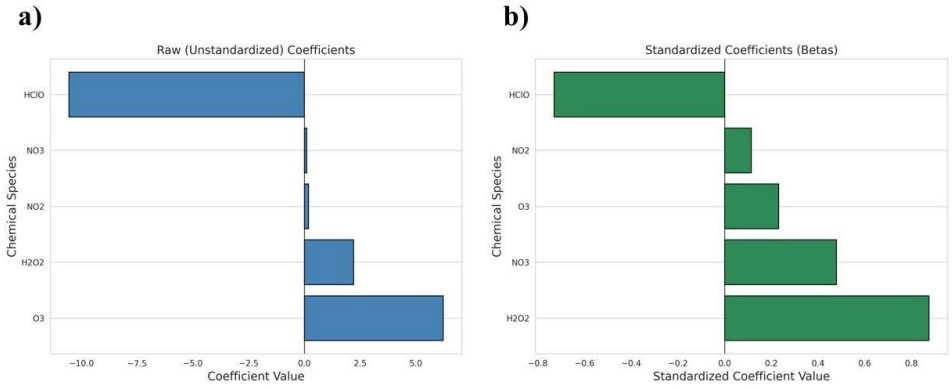

Figure 5. Comparison of raw and standardized regression coefficients for each reactive species, illustrating the influence of predictor scale on coefficient interpretation. (a) Raw (unstandardized) coefficients, expressed in the original units of ORP per mg/L of each species. Because predictors span very different concentration ranges (e.g., O_3: 0.02–2 mg/L; H_2O_2: 9.5–37 mg/L), raw coefficients are not directly comparable across species and can overstate the relative influence of variables with narrower ranges. (b) Standardized coefficients (β), obtained after z-score standardization of all predictors and the response variable, allowing a direct comparison of each species' relative contribution to ORP. Unlike the raw coefficients, which suggested O_3 as the dominant predictor, the standardized coefficients indicate H_2O_2 as the species with the greatest relative statistical influence on ORP, followed by HClO (negative) and NO_3^-.

Residual diagnostics for the regression model are presented in Figure 6. The residuals showed no systematic pattern when plotted against fitted values (**Figure 6a**), and the Shapiro-Wilk test did not indicate a significant deviation from normality (W = 0.89, p = 0.17; **Figure 6b**). The correlation between residuals and activation time was weak (r = -0.09; **Figure 6c**), suggesting that the time-ordered nature of the measurements does not introduce strong autocorrelation into the model residuals, despite the shared temporal trend among several predictors. Cook's distance identified the observations at activation times of 10, 25, and 30 min as exerting disproportionate influence on the fitted model (**Figure 6d**), exceeding the conventional threshold of 4/n = 0.40. This sensitivity to individual observations is an expected consequence of the small sample size and reinforces the need for caution when generalizing the model's coefficients beyond the present dataset.

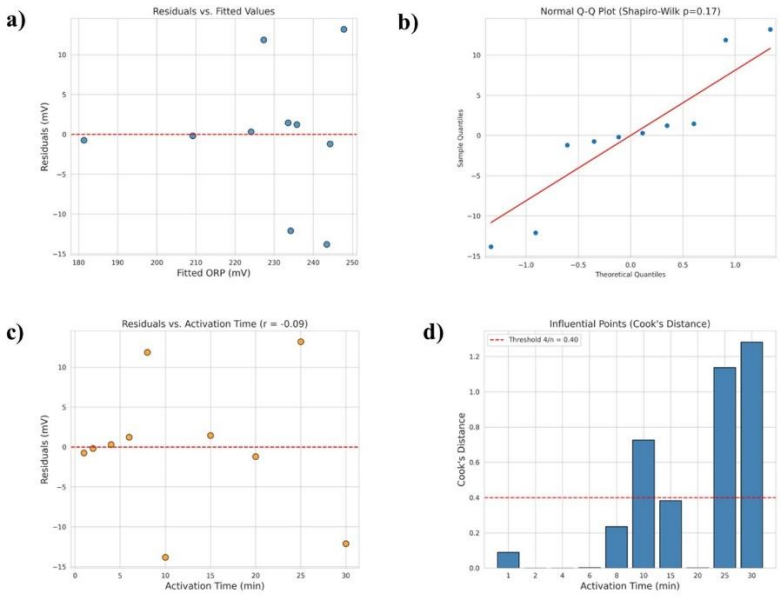

Figure 6. Residual diagnostics for the multiple linear regression model. (a) Residuals plotted against fitted ORP values, used to assess linearity and homoscedasticity; no systematic pattern is observed. (b) Normal Q-Q plot of the residuals, used to assess the normality assumption (Shapiro-Wilk test: W = 0.89, p = 0.17), indicating no significant deviation from normality. (c) Residuals plotted against activation time, used to evaluate potential confounding arising from the time-ordered nature of the measurements (r = -0.09), showing no strong temporal structure in the residuals. (d) Cook's distance for each observation, used to identify influential data points. Observations at activation times of 10, 25, and 30 min exceed the conventional threshold (4/n = 0.40), indicating that these points exert disproportionate influence on the fitted model. This is a relevant limitation given the small sample size (n = 10).

4. Conclusion

Taken together, our results demonstrate that a data-driven, multivariate approach can effectively deconstruct the complex chemistry of plasma-activated saline, while also revealing the practical limits of interpreting such a system from a small experimental dataset. The multiple linear regression model as a whole achieved an in-sample coefficient of determination of $R^2 = 0.84$ (adjusted $R^2 = 0.65$), indicating that the five predictors jointly capture a substantial portion of the variability in ORP within the observed data. However, when evaluated through Leave-One-Out Cross-Validation, the model's predictive coefficient of determination was negative ($Q^2 = -0.33$), indicating that it does not generalize beyond the ten observations used for fitting. The primary value of the model is therefore interpretative and exploratory rather than predictive. The analysis of the model's coefficients revealed strong multicollinearity between key oxidizing species, particularly O_3 and HClO (VIF = 6.66 and 7.32, respectively), and none of the individual coefficients reached conventional statistical significance (all $p > 0.10$).

This statistical interdependency, while complicating the direct interpretation of individual coefficients, is itself an important finding, providing a quantitative signature of the concurrent generation pathways of these species. When raw coefficients were standardized to account for the markedly different concentration ranges of the predictors, H2O2 emerged as the species with the largest relative statistical influence on ORP ($\beta = +0.88$), followed by HClO ($\beta = -0.73$) and NO_3^- ($\beta = +0.48$), while O_3 showed a comparatively modest standardized effect ($\beta = +0.23$) despite its large raw coefficient. This result indicates that the raw-coefficient model's apparent identification of O_3 as the dominant driver was, at least in part, an artifact of its narrow concentration range rather than solely reflecting its individual chemical influence. Considered alongside the electrochemical hierarchy governed by standard reduction potentials ($E°$), O_3 (+2.07 V) remains a chemically plausible contributor to ORP relative to HClO (+1.49 V) and H_2O_2 (+1.78 V); however, given the strong multicollinearity, the absence of individual statistical significance, and the divergence between raw and standardized coefficients, the present dataset does not allow the individual contributions of O_3, H_2O_2,

and HClO to be statistically disentangled with confidence. The negative coefficient for HClO in the raw model is best understood as a mathematical consequence of its collinearity with O_3 rather than as evidence of a specific chemical mechanism.

This work provides a practical case study on the application of a chemometric workflow to complex chemical systems, and, importantly, on the pitfalls that such a workflow must guard against when applied to small, collinear datasets. It demonstrates how standard methods, from EDA to multivariate modeling, are essential not only for prediction but for uncovering underlying data structures such as interdependencies among predictors. At the same time, it shows that diagnostic steps, including variance inflation factors, standardized coefficients, cross-validation, and residual analysis, are indispensable for distinguishing genuine chemical signal from statistical artifacts arising from limited sample size and multicollinearity. This study presents a framework for addressing this challenge by integrating statistical diagnostics with fundamental domain knowledge, rather than relying on model coefficients in isolation.

Ultimately, this analytical framework carries practical implications, albeit more cautious ones than a purely predictive model would suggest. The exploratory evidence indicates that H_2O_2, O_3, and HClO are all statistically and chemically relevant, interdependent contributors to ORP in plasma-activated saline, but does not, on its own, justify prioritizing the optimization of a single species. Reactor and process design aimed at maximizing ORP for therapeutic or sanitizing applications should therefore consider the joint generation of H_2O_2, O_3, and HClO, rather than targeting O_3 only. Confirming the individual contribution of each species, and translating this exploratory analysis into actionable design guidance, will require dedicated experimental designs, such as factorial variation of individual reactive species concentrations, together with larger datasets that can support robust, validated predictive models.

Acknowledgments

Nilton Neto and Felipe Miranda gratefully acknowledge the support from the São Paulo Research Foundation (FAPESP) through individual grants 2023/02268-2 and 2021/14181-3, respectively. The authors also extend their appreciation for the financial assistance provided by FAPESP under Grant Nos. 2019/05856-7 and 2024/1915-0. RSP is a CNPq Universal Program under Grants No. 420775/2023-4.

References:

[1] E. Tsoukou, P. Bourke, and D. Boehm, “Efficacy of plasma activated saline in a co-culture infection control model,” *Sci. Rep.*, vol. 12, no. 1, p. 20230, Nov. 2022, doi: https://doi.org/10.1038/s41598-022-20165-z.

[2] S. Kim and C.-H. Kim, “Applications of Plasma-Activated Liquid in the Medical Field,” *Biomedicines*, vol. 9, no. 11, p. 1700, Nov. 2021, doi: https://doi.org/10.3390/biomedicines9111700.

[3] Y. Jiang *et al.*, “Cold atmospheric plasma-activated saline alleviates secondary injury post-SCI by inhibiting extracellular matrix remodeling and infiltration of proinflammatory macrophages,” *Experimental Neurology*, vol. 383, p. 115004, Jan. 2025, doi: https://doi.org/10.1016/j.expneurol.2024.115004.

[4] M. Ghasemi, H. Najafzadehvarzi, S. Hassani, and A. Ershadian, “Therapeutic potential of nebulized plasma activated saline (N-PAS) in treating pulmonary infections in rats,” *Microbial Pathogenesis*, vol. 206, p. 107699, Sep. 2025, doi: https://doi.org/10.1016/j.micpath.2025.107699.

[5] V. Veronico *et al.*, “Anticancer Effects of Plasma-Treated Water Solutions from Clinically Approved Infusion Liquids Supplemented with Organic Molecules,” *ACS Omega*, vol. 8, no. 37, pp. 33723–33736, 2023, doi: https://doi.org/10.1021/acsomega.3c04061.

[6] V. Jirásek and P. Lukeš, “Formation of reactive chlorine species in saline solution treated by non-equilibrium atmospheric pressure He/O2 plasma jet,” *Plasma Sources Sci. Technol.*, vol. 28, no. 3, 2019, doi: https://doi.org/10.1088/1361-6595/ab0930.

[7] A. Mai-Prochnow *et al.*, “Interactions of plasma-activated water with biofilms: inactivation, dispersal effects and mechanisms of action,” *NPJ Biofilms Microbiomes*, vol. 7, no. 1, p. 11, Jan. 2021, doi: https://doi.org/10.1038/s41522-020-00180-6.

[8] A. C. Borges *et al.*, "Applications of Cold Atmospheric Pressure Plasma in Dentistry," *Applied Sciences*, vol. 11, no. 5, p. 1975, Feb. 2021, doi: https://doi.org/10.3390/app11051975.

[9] T. F. de Melo *et al.*, "Plasma–Saline Water Interaction: A Systematic Review," *Materials*, vol. 15, no. 14, p. 4854, Jul. 2022, doi: https://doi.org/10.3390/ma15144854.

[10] S. Xu *et al.*, "Anticancer effects of DBD plasma-activated saline within different discharge modes," *J. Phys. D Appl. Phys.*, vol. 56, no. 34, p. 345205, Aug. 2023, doi: https://doi.org/10.1088/1361-6463/acd0bb.

[11] N. F. Azevedo Neto *et al.*, "Serial association of dielectric barrier discharge and gliding arc systems: a novel approach for plasma-activated saline production," *J. Phys. D Appl. Phys.*, vol. 58, no. 1, p. 015202, Jan. 2025, doi: https://doi.org/10.1088/1361-6463/ad800e.

[12] N. F Azevedo Neto, F. S. Miranda, P. W. P Moreira Junior, and R. S. Pessoa, "Influence of gas composition on reactive species generation in plasma-activated saline using two-stage plasma reactors system," *Phys. Scr.*, vol. 100, no. 7, 2025, doi: https://doi.org/10.1088/1402-4896/adebb9.

[13] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using python," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4727–4735, 2019, doi: https://doi.org/10.35940/ijitee.L3591.1081219.

[14] A. Bezerra, I. Silva, L. A. Guedes, D. Silva, G. Leitão, and K. Saito, "Extracting value from industrial alarms and events: A data-driven approach based on exploratory data analysis," *Sensors (Switzerland)*, vol. 19, no. 12, 2019, doi: https://doi.org/10.3390/s19122772.

[15] J. H. Kalivas, "Expanding the Chemometric Data Analysis Toolbox With Immersive Analytics," *Journal of Chemometrics*, vol. 39, no. 9, pp. 1–5, Sep. 2025, doi: https://doi.org/10.1002/cem.70060.

[16] T. S. Madssen, A. Smilde, J. Camacho, A. H. Jarmund, J. Westerhuis, and G. F. Giskeødegård, "Statistical Validation of Multivariate Treatment Effects in Longitudinal Study Designs," *Journal of Chemometrics*, vol. 39, no. 8, Aug. 2025, doi: https://doi.org/10.1002/cem.70044.

[17] P. C. Austin and E. W. Steyerberg, "The number of subjects per variable required in linear regression analyses," *J. Clin. Epidemiol.*, vol. 68, no. 6, pp. 627–636, Jun. 2015, doi: https://doi.org/10.1016/j.jclinepi.2014.12.014.

[18] M. L. Waskom, "seaborn : statistical data visualization Statement of need," vol. 6, pp. 1–4, 2021, doi: https://doi.org/10.21105/joss.03021.

[19] J. Hao and T. K. Ho, "Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language," *Journal of Educational and Behavioral Statistics*, vol. 44, no. 3, pp. 348–361, Jun. 2019, doi: https://doi.org/10.3102/1076998619832248.

[20] F. Nelli, *Python Data Analytics*, 2nd ed. Berkeley, CA: Apress, 2018. doi: https://doi.org/10.1007/978-1-4842-3913-1.

[21] R. M. O'brien, "A Caution Regarding Rules of Thumb for Variance Inflation Factors," *Qual. Quant.*, vol. 41, no. 5, pp. 673–690, Sep. 2007, doi: https://doi.org/10.1007/s11135-006-9018-6.

[22] P. Filzmoser, B. Liebmann, and K. Varmuza, "Repeated double cross validation," *J. Chemom.*, vol. 23, no. 4, pp. 160–171, Apr. 2009, doi: https://doi.org/10.1002/cem.1225.

[23] H. J. Kim *et al.*, "Analyses of the Chemical Composition of Plasma-Activated Water and Its Potential Applications for Vaginal Health," *Biomedicines*, vol. 11, no. 12, 2023, doi: https://doi.org/10.3390/biomedicines11123121.

[24] J. Ma *et al.*, "Unusual Hypochlorous Acid (HClO) Recognition Mechanism Based on Chlorine–Oxygen Bond (Cl–O) Formation," *Chemistry – A European Journal*, vol. 25, no. 29, pp. 7168–7176, May 2019, doi: https://doi.org/10.1002/chem.201806264.

[25] K. M. Yeater, S. E. Duke, and W. E. Riedell, "Multivariate Analysis: Greater Insights into Complex Systems," *Agronomy Journal*, vol. 107, no. 2, pp. 799–810, Mar. 2015, doi: https://doi.org/10.2134/agronj14.0017.

[26] Z. C. Liu, L. Guo, D. X. Liu, M. Z. Rong, H. L. Chen, and M. G. Kong, "Chemical Kinetics and Reactive Species in Normal Saline Activated by a Surface Air Discharge," *Plasma Processes and Polymers*, vol. 14, no. 4–5, pp. 1–9, 2017, doi: https://doi.org/10.1002/ppap.201600113.

[27] M. S. Siddiqui, "Chlorine-ozone interactions: Formation of chlorate," *Water Research*, vol. 30, no. 9, pp. 2160–2170, Sep. 1996, doi: https://doi.org/10.1016/0043-1354(96)00071-1.

[28] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970, doi: https://doi.org/10.1080/00401706.1970.10488634.

[29] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, Oct. 2001, doi: https://doi.org/10.1016/S0169-7439(01)00155-1.

[30] A. J. Bard, *Standard Potentials in Aqueous Solution*, 1st ed. New York: Routledge, 2017. doi: https://doi.org/10.1201/9780203738764.