engineering proceedings

MDPI

Proceeding Paper

Application of Machine Learning for the Prediction of Coulombic Efficiency in Lithium Metal Batteries †

Sergio Rubén Ocampo-Pérez [1,*], Noureddine Lakouari [2,3] and Outmane Oubram [4,*]

[1] Maestría en Optimización y Cómputo Aplicado, Facultad de Contaduría, Administración e Informática (FCAeI), Universidad Autónoma del Estado de Morelos, Cuernavaca 62200, Morelos, Mexico
[2] Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Insurgentes Sur 1582, Ciudad de México 03940, Mexico; n.lakouari@inaoep.mx
[3] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), San Andrés Cholula 72840, Puebla, Mexico
[4] Facultad de Ciencias Químicas e Ingeniería, Universidad Autónoma del Estado de Morelos (UAEM), Cuernavaca 62210, Morelos, Mexico
* Correspondence: sergio.ocampo@fcaei.uaem.edu.mx (S.R.O.-P.); oubram@uaem.mx (O.O.)
† Presented at the 2nd International Conference on Sciences and Techniques for Renewable Energy and the Environment, Al Hoceima, Morocco, 28–30 April 2026.

Abstract

The commercialization of lithium metal batteries, a key technology for high-density energy storage, is hindered by issues with coulombic efficiency, which dictates battery stability and life. In this paper, we propose a machine learning framework to forecast liquid electrolyte efficiency, where two experimental data sources were combined to create a curated dataset of 283 records. In addition, to assess several ensemble learning algorithms, thirteen chemical descriptors were used, as well as interpretability analysis and Bayesian optimization to guarantee physicochemical consistency. We found that the optimized CatBoost model achieved a coefficient of determination (R2) of 0.61 on the test set and a mean squared error (MSE) of 0.0924, representing a significant improvement in predictive accuracy compared to previous standards. Furthermore, these results demonstrate that regulating oxygen levels in solvent environments is a key component of high-density energy storage. These results can serve as a virtual screening tool in order to discover high-performance electrolytes with the minimum experimental costs.

Keywords: Lithium metal batteries; coulombic efficiency; machine learning; CatBoost; XGBoost; SHAP interpretability; electrolyte design

1. Introduction

Nowadays, many electronic applications depend on lithium-ion batteries due to their lower cost and high energy density, which allows their use in portable devices to large-scale energy storage systems. However, one of the main challenges associated with these batteries is increasing their coulombic efficiency (CE), which measures the reversibility of the electrochemical system over repeated charge–discharge cycles. Achieving values greater than 99.9% in CE is critical to ensure high capacity retention beyond 1000 cycles [1]. On the other hand, lower CE often indicates parasitic side reactions and lithium dendrite formation, which means a degraded battery lifetime. Recent studies suggest that in high-performing electrolytes, galvanic corrosion at the lithium–current collector interface acts as a hidden driver of efficiency loss, preventing the system from surpassing the 99.9% threshold [2].

Machine learning can be used to understand and predict those mechanisms based on reliable experimental data. However, measuring coulombic efficiency in liquid electrolytes

Check for updates

Academic Editors: Charaf Laghlimi, Younes Ziat, Zakaryaa Zarhri and Hamza Belkhanchi

Published: 23 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

remains a complex and expensive task because they require the fabrication of complete battery cells and extensive charge–discharge cycles under controlled conditions. Furthermore, across different studies, the lack of clear standardization in measurement and reporting protocols hinders the comparability and reproducibility of results. Therefore, supervised machine learning models combined with hyperparameter optimization techniques have become a powerful approach to bypassing traditional experimental design. This methodology has been validated by recent research demonstrating machine learning's ability to link microscopic properties to overall performance in high-voltage cells. For example, it has accelerated the discovery of superelectrolytes across a wide range of designs by identifying critical chemical indicators that enhance capacity retention [3]. Ref. [4] demonstrated that a data-driven approach could identify the reduction in solvent oxygen content as a primary driver for high CE. Ref. [5] introduced the 'Electrolytomics' framework, which, while centered on ionic conductivity, also demonstrates the capability of ML to target multiple properties including electrochemical stability and CE through a unified big data approach. Therefore, integrating predictive algorithms reduces reliance on trial-and-error methods and improves the accuracy of composition selection.

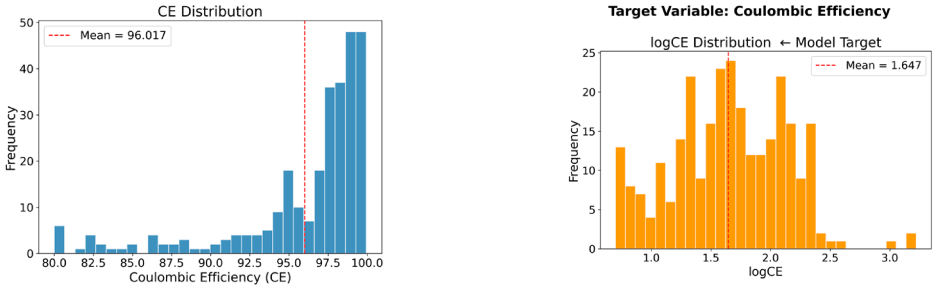
Even with these advancements, the performance of machine learning models in this area remains limited due to the diversity of experimental data and its scarcity. In this paper, a structured dataset was generated by combining two sources to expand the chemical range and enhance the predictive power of Coulombic efficiency regression models. This approach aims to reduce prediction error and enable more reliable predictions, particularly in the high-efficiency range. The rest of the research paper was organized as follows: Section 2 is dedicated to the methodology and dataset, Section 3 is dedicated to the results, Section 4 is dedicated to the discussion, and the last section is dedicated to the conclusions.

2. Materials and Methods

2.1. Dataset Collection and Curation

The initial database was constructed by integrating 150 samples from [4], and 142 samples from [5]. The integration of these heterogeneous sources was justified by their high degree of experimental alignment; specifically, both datasets [4,5] utilized standardized Li | Cu coin cell configurations and restricted their measurements to a narrow current density range of 0.4 to 1.0 mA cm^{-2} at room temperature. During the data curation process, 9 records from the [5] dataset were identified as outliers and removed. These samples exhibited Coulombic Efficiency (CE) values below 80% (reaching as low as 9%), which is significantly outside the high-performance regime analyzed in this study. The decision to exclude these records was based on the premise that such extreme values typically indicate catastrophic cell failure, internal short circuits, or atypical experimental artifacts that do not represent the inherent electrochemical stability of the electrolyte. Including these non-representative data points would introduce significant noise, preventing the model from accurately capturing the subtle physicochemical interactions that govern the high-efficiency window (>98%), which is the primary focus of practical battery development. This consistency in cell architecture and operational parameters ensures that the target variable reflects the intrinsic chemical stability of the electrolytes across both sub-datasets, effectively mitigating concerns regarding systematic bias or domain shift. As a result, there are 283 records in the final curated dataset. The distribution of CE is mostly concentrated in the high-efficiency region (>95%), which is crucial for realistic lithium metal battery applications, as seen in Figure 1. To guarantee consistency, Table 1 provides comprehensive descriptive statistics and ranges for the 13 chemical descriptors.

https://doi.org/10.3390/engproc2026144003



(**a**) (**b**)

Figure 1. Coulombic Efficiency (CE) distribution: (**a**) percentage scale; (**b**) log-scale (LCE).

Table 1. Descriptive statistics and ranges for all 13 chemical descriptors (N = 283).

Category	Abbr.	Min	Max	Mean	Std.
Total Molar Fraction	O	0.026	0.334	0.185	0.065
	C	0.132	0.350	0.244	0.032
	F	0.000	0.362	0.097	0.086
Solvent Environment	sO	0.000	0.283	0.149	0.068
	sC	0.121	0.350	0.241	0.034
	sF	0.000	0.338	0.060	0.091
Anion Environment	aO	0.000	0.182	0.036	0.039
	aC	0.000	0.058	0.003	0.008
	aF	0.000	0.183	0.036	0.026
Chemical Ratios	FO	0.000	7.746	0.697	0.913
	FC	0.000	1.437	0.410	0.363
	OC	0.115	2.122	0.779	0.325
	InOr	0.321	3.526	1.355	0.591

2.2. Target Variable: Logarithmic Coulombic Efficiency

We chose the logarithmic coulombic efficiency (LCE) as the target variable for regression, as this transformation effectively captures changes near the convergent limit of coulombic efficiency approaching 100%, as described in Figure 1, which are critically important for battery lifetime [4]. Where the mathematical relationship of LCE is defined as:

$$LCE = -log10(1 - CE/100) \quad (1)$$

2.3. Selection of Chemical Descriptors

Thirteen independent variables were selected for the electrolyte representation, following the elemental-based framework established by [4]. These descriptors are divided into four groups based on their chemical makeup and function: Total Molar Fraction, Solvent Environment, Anion Environment, and Chemical Ratios. This classification is grounded in the fact that elemental distribution serves as a proxy for critical physicochemical properties, such as Li-ion solvation structures and the resulting Solid Electrolyte Interphase (SEI) composition. To evaluate the statistical relationship between these features, a pairwise correlation analysis was performed (see Appendix A, Figure A1). Although high correlations are observed between total elemental fractions and their respective solvent environments (e.g., C and sC, r = 0.97), these descriptors are maintained to ensure a comprehensive representation of both the bulk and localized chemical environments. We provide the precise range (minimum and maximum), mean, and standard deviation for every descriptor over the 283 samples in Table 1. These statistical parameters guarantee the prediction framework's transparency and repeatability by defining the chemical space of the investigation,

enabling the model to capture the complex interactions of additives and solvents without the need for expensive first-principles simulations.

2.4. Regression Model Architecture

A supervised learning workflow was implemented by comparing several ensemble learning algorithms to identify the architecture with the lowest generalization error. The evaluated models included XGBoostRegressor, Extra Trees, LGBMRegressor, and CatBoost. These gradient boosting and randomized tree-based methods are widely used for regression problems due to their ability to model nonlinear relationships and complex feature interactions.

Training Configuration and Optimization with Optuna

The dataset was randomly partitioned into training (70%), validation (15%), and test (15%) subsets using a fixed random seed (random state = 42) to ensure reproducibility. Given the continuous nature of the logarithmic Coulombic efficiency (LCE) target variable, stratification was not applied. Nevertheless, the distribution of LCE values was carefully examined across all subsets (see Appendix A, Table A1) to confirm that they retained comparable statistical characteristics. Because the dataset consists of independent chemical compounds, no specific measures were required to mitigate data leakage.

Hyperparameter optimization was conducted using the Optuna framework, which enables flexible and efficient exploration of the search space through adaptive sampling strategies [6]. A Bayesian optimization approach was employed to minimize the mean squared error (MSE) on the validation set, allowing each model to be fine-tuned to its optimal configuration before final evaluation. To further evaluate model robustness and account for the uncertainty inherent in the small dataset size, a bootstrap resampling analysis with 1000 iterations was performed. This approach provides a more rigorous estimation of the confidence intervals for the performance metrics, especially given the limited number of samples in the test set.

3. Results

3.1. Model Performance Evaluation and Optimal Model Selection

The generalization capability of the implemented architectures was evaluated by comparing their performance across the training, validation, and test sets. A relatively small performance gap between these three groups was observed for all evaluated models, indicating that the hyperparameter optimization performed using Optuna (version 4.7.0) was effective in preventing overfitting.

The consistency of the results across the different data splits suggests that the models capture meaningful physicochemical relationships in the electrolyte compositions rather than relying on simple memorization of the training data. The quantitative comparison of the models, measured using the coefficient of determination (R^2), is summarized in Table 2.

To evaluate the statistical significance of the CatBoost results, a bootstrap analysis was performed on the test set. As shown in Table 2, the model maintains a mean R2 of 0.59 within a 95% confidence interval of [0.40, 0.72], providing a robust estimate of performance uncertainty.

After confirming the stability of the evaluated models, the optimal algorithm was selected based on its predictive performance on the independent test set. The CatBoost model emerged as the most accurate tool for the regression task, achieving a coefficient of determination of R^2 = 0.6097 and the lowest MSE = 0.0924. The small gap between training (R2 = 0.6185) and test (R2 = 0.6097) performance suggests that the Bayesian optimization and regularization parameters (L2 leaf regularization) effectively prevented

overfitting, allowing the model to generalize well despite the small dataset size. The optimal hyperparameters obtained through Optuna are summarized in Table 3.

Table 2. Comparison of model performance measured using the *R*2 across the training, validation, and test datasets.

Model	R2				MSE		
	Train	Validation	Test	Bootstrap Mean R2 [95% CI] *	Train	Validation	Test
XGBoost	0.6380	0.5906	0.5333	0.5067 [0.2740, 0.6487]	0.0872	0.0717	0.1127
Extra Trees	0.6054	0.4933	0.5853	0.5796 [0.4033, 0.7059]	0.0907	0.0915	0.0993
CatBoost	0.6185	0.5064	0.6097	0.5916 [0.4053, 0.7289]	0.0893	0.0979	0.1078
LightGBM	0.6118	0.4508	0.5482	0.5342 [0.3821, 0.6469]	0.0875	0.0871	0.0918

* Bootstrap results were obtained using n = 1000 iterations. The reported value is the mean R2 followed by the 95% confidence interval in brackets.

Table 3. Final optimized hyperparameters for the CatBoost model via Optuna.

Hyperparameter	Value	Description
iterations	436	Number of boosting rounds (trees).
depth	2	Depth of the trees (complexity control).
learning_rate	0.01016	Step size shrinkage to prevent overfitting.
l2_leaf_reg	0.03706	L2 regularization coefficient for leaf values.
subsample	0.59682	Fraction of samples used for bagging.
colsample_bylevel	0.52163	Fraction of features used for each split level.

This superior performance suggests that the CatBoost architecture more effectively captures the nonlinear relationships inherent to the compositional and structural descriptors, enabling a more precise prediction of LCE compared with the other evaluated ensemble techniques.

Figure 2 presents a quantitative comparison of the Mean Squared Error (MSE) across the evaluated architectures. The CatBoost model achieved a notably lower error (0.0918) compared to the XGBoost baseline (0.1127), confirming that its gradient boosting implementation is better suited for capturing the nonlinear interactions of these chemical descriptors.

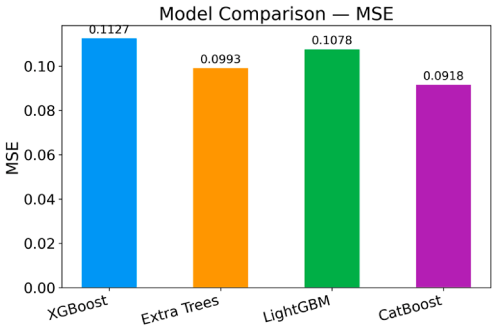
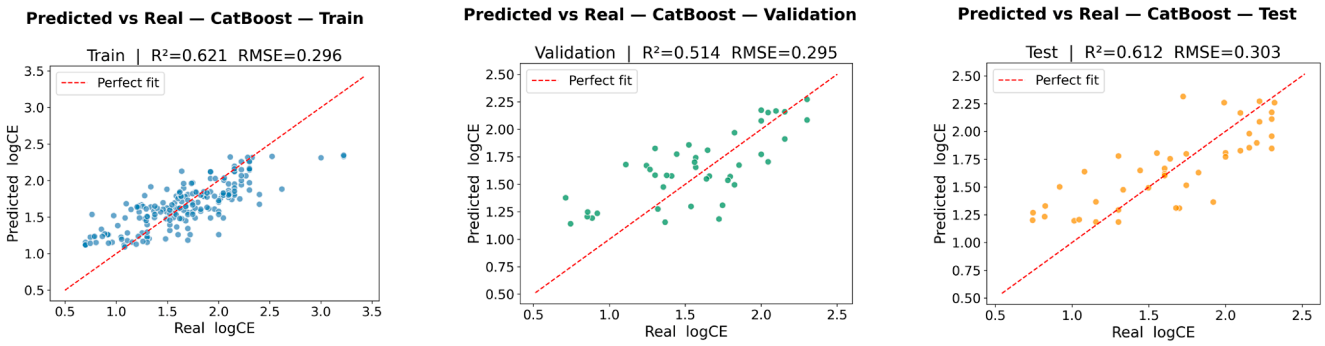


Figure 2. Presents a quantitative comparison of the Mean Squared Error (MSE) across the evaluated architectures. The error metrics are reported based on the Logarithmic Coulombic Efficiency (LCE) scale used during model training.

To test how well the best model predicts, we compare the calculated LCE values against the actual results using scatter plots. Figure 3 shows the predicted versus observed values for the CatBoost model across the training, validation, and test datasets. Data points close to the diagonal line indicate good agreement and reliable model predictions.



(**a**) (**b**) (**c**)

Figure 3. Predicted versus experimental values of LCE obtained with the CatBoost model. (**a**) Training set; (**b**) Validation set; (**c**) Test set. The dashed line represents the ideal $y = x$ relationship.

3.2. Interpretation of the Model Using SHAP Values

To understand the logic behind the CatBoost model's predictions and validate its physicochemical consistency, a SHAP-based analysis was employed. This approach allows identification of which descriptors drive the LCE increase and which act as limiting factors.

Global Feature Importance

Figure 4 depicts a summary plot of SHAP, where we can see an interpretation of how each descriptor contributes. We found that both elemental composition and descriptors related to the chemical environment play a great role in the model predictions.

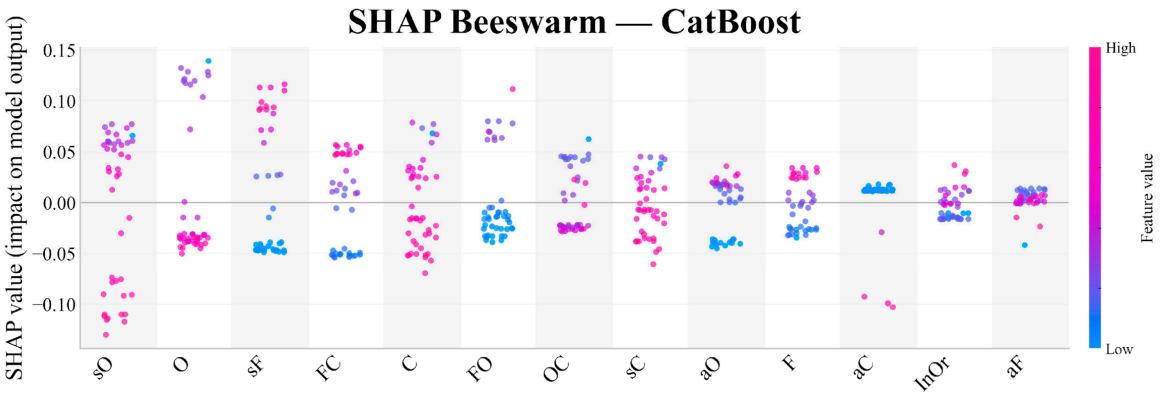


Figure 4. SHAP summary plot showing the global importance of descriptors in the CatBoost model for LCE prediction.

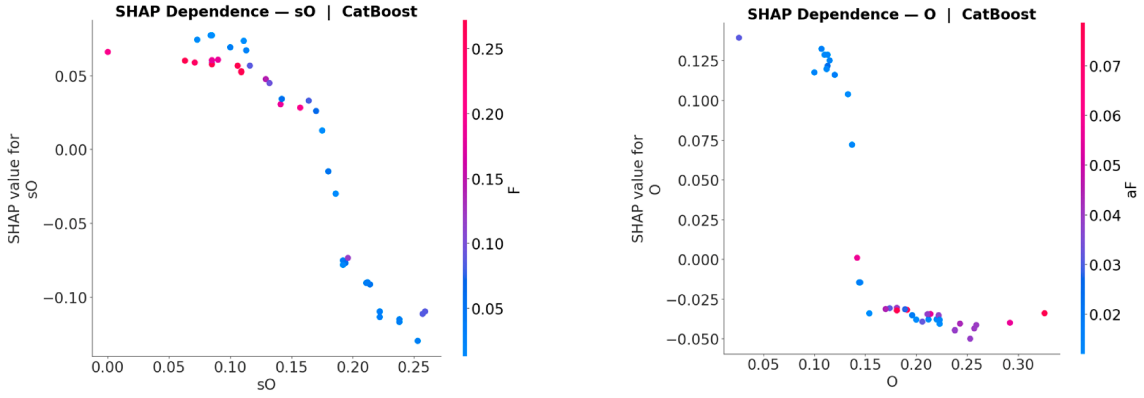
Hence, the sO is the most influential descriptor, where a clear trend is observed in which lower sO values contribute positively to the predicted LCE, while higher values negatively impact the model output. This behavior is mechanistically linked to the high polarity of carbon–oxygen bonds in traditional solvents, which facilitates parasitic reduction reactions and prevents the formation of a stable passivation layer [1]. Our findings align with the observations reported by [4], where they found that reducing oxygen-rich solvent environments enhances the stability of lithium metal anodes and mitigates reactions.

The chemical environment of the solvent and the balance between oxygenated and fluorinated species play a key role in modifying the performance of the electrolyte, and this is clearly evident from the fact that the variables related to elemental composition, such as O, C, and fluorine-containing descriptors (sF, FO), show a systemic effect on the predictions.

https://doi.org/10.3390/engproc2026144003

From this, it is clear that Coulomb efficiency is governed by a complex interaction between the solvent environment and the elemental composition, and that, to achieve high efficiency in lithium metal electrolytes, controlling the oxygen content and incorporating fluorine-rich components are crucial.

Figure 5a shows that when sO is low, the predicted LCE values increase, but once sO exceeds 0.15, its effect becomes negative, penalizing the LCE. The red points, which indicate a high fluorine content, tend to cluster in regions with low oxygen levels. This suggests a synergistic effect, where maximizing LCE not only requires low oxygen content in the solvent, but is further enhanced when the electrolyte is rich in fluorine. Quantitatively, this synergy reflects the higher binding energy and easier donation of fluorine atoms compared to oxygenated moieties, which prioritizes the formation of a robust inorganic SEI over more fragile organic-oxygenated phases [7]. Figure 5b shows that when total oxygen O is less than 0.12, the LCE increases, but as soon as this range is exceeded, the value becomes negative and the substance's performance drops, even when the amount of fluorine is moderate (purple dots), highlighting the amount of O as a decisive variable for high performance.



(**a**) (**b**)

Figure 5. SHAP dependence plots for oxygen features. (**a**) Molar fraction of oxygen in solvents (sO) interacting with fluorine (F); (**b**) Total molar fraction of oxygen (O) interacting with fluorine in anions (aF).

4. Discussion

The use of a logarithmic transformation ensures a consistent comparison framework. The CatBoost model achieved a Mean Squared Error (MSE) of 0.0924, which corresponds to a Root Mean Squared Error (RMSE) of 0.304 (RMSE = $\sqrt{\text{MSE}}$), as shown in the scatter plots in Figure 3. This performance represents a 73% reduction in MSE compared to the literature baseline of 0.343. The error sensitivity analysis in Figure 6 highlights the model's reliability. Unlike the literature baseline [4], our CatBoost model maintains tighter confidence intervals and lower variance in the high-efficiency region (>99% CE), which is the most critical zone for the practical commercialization of lithium metal batteries. Despite the model's accuracy, it is important to acknowledge that the dataset of 283 records is relatively small, which may limit generalizability to entirely new classes of electrolytes not represented in the training data.

The use of SHAP analysis further strengthens this work by providing a window into the model's 'black box,' confirming that the predictions are grounded in established electrochemical principles rather than mere statistical correlation.

https://doi.org/10.3390/engproc2026144003

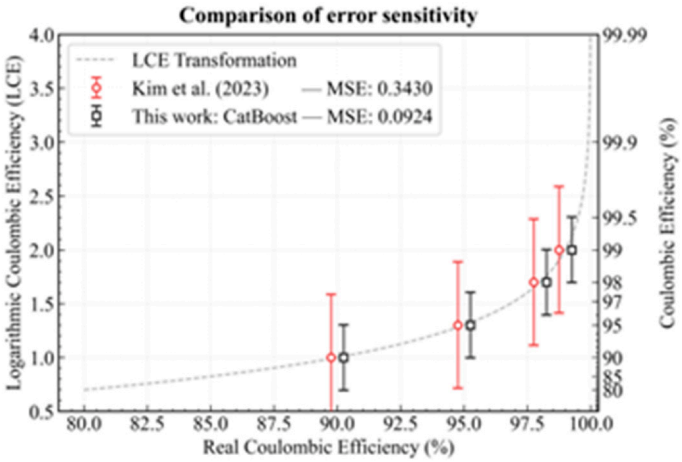


Figure 6. Comparison of sensitivity to error between CatBoost and the *LCE* literature values from Kim et al. (2023) [4].

5. Conclusions

In this paper, we found that the CatBoost model, optimized using the Optuna library, is the best model for predicting the logarithmic coefficient of lithium metal batteries. The model performed exceptionally well on the test set, with a coefficient of determination R^2 of 0.6097 and a mean squared error of 0.0924.

We expanded the scope of the chemical analysis to include 283 samples by combining datasets from the references [4,5], resulting in a significant 73% reduction in prediction error compared to the baseline results. To better understand the model's behavior, we used SHAP analysis. This analysis revealed that the solvent oxygen (sO) content is the most influential factor, confirming that low-oxygen environments play a crucial role in the stability of the lithium anode.

Overall, the model demonstrated high reliability, particularly in high-efficiency systems (over 98%). This makes it a powerful virtual testing tool that can help guide the design of advanced electrolytes, while reducing the need for costly and time- consuming experiments.

Author Contributions: Conceptualization, S.R.O.-P., N.L. and O.O.; methodology, S.R.O.-P., N.L. and O.O.; software, S.R.O.-P., N.L. and O.O.; validation, S.R.O.-P., N.L. and O.O.; formal analysis, S.R.O.-P., N.L. and O.O.; investigation, S.R.O.-P., N.L. and O.O.; resources, S.R.O.-P., N.L. and O.O.; data curation, S.R.O.-P., N.L. and O.O.; writing—original draft preparation, S.R.O.-P., N.L. and O.O.; writing—review and editing, S.R.O.-P., N.L. and O.O.; visualization, S.R.O.-P., N.L. and O.O.; supervision, N.L. and O.O.; project administration, S.R.O.-P., N.L. and O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code developed for this study is openly available in the following GitHub repository: https://github.com/sergioruben/Application-of-ML-for-the-Prediction-of-Coulombic-Efficiency-in-Lithium-Metal-Batteries/ (accessed on 20 May 2026). The original experimental datasets used in this study were obtained from Kim et al. (2023) [4] and Kumar et al. (2025) [5]. As the datasets are the intellectual property of their respective authors they cannot be redistributed directly. Readers are directed to the original publications for data access.

Acknowledgments: We would like to thank the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), formerly CONAHCyT, for funding the postgraduate studies linked to this research with a full scholarship. We are also very grateful to the Universidad Autónoma del Estado de Morelos (UAEM) and the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) for all their institutional support, facilities, and guidance throughout this project.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CE	Coulombic Efficiency
LCE	Logarithmic Coulombic Efficiency
ML	Machine Learning
MSE	Mean Squared Error
R^2	Coefficient of Determination
SHAP	SHapley Additive exPlanations
sO	Molar fraction of Oxygen in solvents
sF	Molar fraction of Fluorine in solvents
InOr	Inorganic/Organic ratio

Appendix A

Table A1. Descriptive statistics of the Logarithmic Coulombic Efficiency (LCE) target variable across the training, validation, test, and full datasets.

Metric	Training (70%)	Validation (15%)	Test (15%)
Sample size (n)	198	42	43
Mean	1.6643	1.5669	1.6436
Std. Deviation	0.4814	0.4286	0.4924
Minimum	0.699	0.7122	0.7423
Median (Q2)	1.6646	1.5686	1.6778
Maximum	3.2218	2.301	2.3188
Skewness	0.127	−0.2591	−0.2749

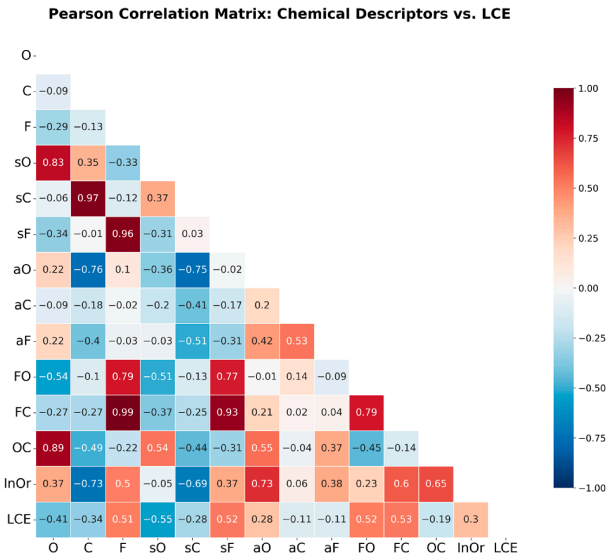


Figure A1. Pairwise correlation heatmap of the thirteen chemical descriptors and the logarithmic Coulombic efficiency target variable. The numerical values denote Pearson correlation coefficients, with the color intensity representing the strength of the linear relationship between the electrolyte's elemental distribution and the model's predictive target.

References

1. Hobold, G.M.; Lopez, J.; Guo, R.; Minafra, N.; Banerjee, A.; Shirley Meng, Y.; Shao-Horn, Y.; Gallant, B.M. Moving beyond 99.9% Coulombic efficiency for lithium anodes in liquid electrolytes. *Nat. Energy* **2021**, *6*, 951–960. [CrossRef]
2. Oyakhire, S.T.; Kim, S.C.; Zhang, W.; Shuchi, S.B.; Cui, Y.; Bent, S.F. Galvanic corrosion underlies coulombic efficiency differences in high-performing lithium metal battery electrolytes. *Energy Environ. Sci.* **2025**, *18*, 4847–4858. [CrossRef]
3. Yan, Y.; Hai, F.; Wang, B.; Cao, W.; Li, M.; Wang, C.; Li, N.; Zhao, D. Machine learning accelerates high-voltage electrolyte discovery for lithium metal batteries. *Energy Storage Mater.* **2025**, *79*, 104312. [CrossRef]
4. Kim, S.C.; Oyakhire, S.T.; Athanitis, C.; Wang, J.; Zhang, Z.; Zhang, W.; Boyle, D.T.; Kim, M.S.; Yu, Z.; Gao, X.; et al. Data-driven electrolyte design for lithium metal anodes. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2214357120. [CrossRef] [PubMed]
5. Kumar, R.; Vu, M.C.; Ma, P.; Amanchukwu, C.V. Electrolytomics: A Unified Big Data Approach for Electrolyte Design and Discovery. *Chem. Mater.* **2025**, *37*, 2720–2734. [CrossRef]
6. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019. [CrossRef]
7. Xue, W.; Shi, Z.; Huang, M.; Feng, S.; Wang, C.; Wang, F.; Lopez, J.; Qiao, B.; Xu, G.; Zhang, W.; et al. FSI-inspired solvent and "full fluorosulfonyl" electrolyte for 4 V class lithium-metal batteries. *Energy Environ. Sci.* **2020**, *13*, 212–220. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.