Original Research Article

Comparing Methods of Synthetic Data Generation and Determining Their Feasibility in Lithium-Ion Battery Research

Jayden Yang

Walnut High School, 400 Pierre Road, Walnut, CA 91789, United States

ABSTRACT

Lithium-ion batteries (LIBs) have gradually developed into a cornerstone of modern renewable energy systems over the past few decades. However, research efforts into LIBs have been severely limited by the lack of public LIB data, creating a developmental bottleneck in this field. In order to address this prominent global issue, this study develops and compares two contrasting synthetic generators, physics-aware (PADS) and distributional (DDS), to determine each method's feasibility in real-world LIB research. Using both methods, the study first synthetically expands an existing public LIB dataset. Each synthetically generated dataset is then statistically compared to the original data through a custom multilevel evaluation framework to assess which approach better preserves statistical fidelity. Ultimately, this synthetic method comparison study aims to outline a superior model that can be utilized for data expansion in real-world LIB research. When writing this paper, it was initially hypothesized that the distributional synthetic data generation method would better preserve statistical fidelity than the physics-aware synthetic data generation method. Indeed, after cycling through the research process for both synthetic data generation methods, the results were generally consistent with the initial claims. However, deeper analysis into the synthetic data quality revealed instability in the regression relationships and substantial diagnostic limitations in both models, highlighting a significant limitation in their application. This study concludes that while distributional data generators are more well-suited for statistical fidelity and general LIB research, the utilization of each model relies heavily on its intended application. Therefore, the results of this research can be used as a foundation for further comparative investigations between synthetic data generation methods in the LIB research field.

Keywords: Lithium-Ion Batteries; Synthetic Data Generation; Data Augmentation; Regression Modeling; Energy Storage

INTRODUCTION

The global transition towards a world powered by renewable energy currently faces a major limitation: intermittency. The biggest problem with natural energy production is that its supply doesn't always match real-time demand, creating instability (1). As a result, the transition to renewables depends heavily on reliable energy storage systems that can balance supply and demand. As of today, one of the dominant choices for energy storage technologies has been lithium-ion batteries (LIBs) due to their high energy density, long cycle life, and efficiency. LIBs have already developed into an integral part of everyday life; electric vehicles,

Corresponding author: Jayden Yang, E-mail: jjaydenyang@gmail.com.

Accepted August 19, 2026
https://doi.org/10.70251/HYJR2348.4412331254

smartphones, tablets, and even large-scale grid storage facilities are all supported by them (2). This fact, along with the optimal characteristics they possess, positions LIBs as a cornerstone of modern sustainability developments around the world. Therefore, improving the performance and reliability of LIBs is imperative for accelerating progression towards a world powered by low-carbon energy systems. However, there remains a massive roadblock for the LIB research community: due to their expensive nature, the current available data on LIBs is insufficient for meaningful scientific research, making data availability a major bottleneck for scientific progress (3).

The tradeoff between physical plausibility and statistical fidelity is particularly important in LIB research, since battery analysis requires both statistical fidelity and physical plausibility. However, while both of these metrics are essential for high-quality synthetic data, statistical fidelity is a foundational requirement, as any dataset that is unable to replicate basic structural relationships and quantities cannot be meaningfully applied to modeling or data analysis (4). Therefore, this study will only focus on evaluating statistical fidelity as a necessary attribute of the dataset, while acknowledging that physical plausibility is another important component of LIB data that will not be addressed in this paper.

This study aims to address this problem by investigating the feasibility of two distinct synthetic dataset generation methods that artificially expand an existing LIB dataset while also maintaining statistical fidelity. To complement this, a paper-specific process has been created to analyze the effectiveness of each synthetic data generation method and determine whether or not the strategy is truly feasible for real-world research purposes. The research system is structured to assess synthetic data generation in battery science through a model-specific framework. This study will ultimately provide clarification on both the benefits and limitations of different synthetic options, and then determine which approach more effectively preserves statistical fidelity when expanding an existing LIB dataset. Increased data availability enhances the accuracy of battery modeling, which would ultimately contribute to more reliable and efficient energy storage systems worldwide (5).

LITERATURE REVIEW

Key Concepts and Evaluation Criteria

Several related terms are used throughout this study and are defined here to maintain consistency. Statistical fidelity refers to how well a synthetic dataset preserves important statistical properties of the original dataset. In this study, statistical fidelity includes similarities in univariate descriptive statistics, pairwise correlation structures, and fitted multivariable regression relationships. However, statistical fidelity does not mean that synthetic observations are experimentally measured or that they reproduce every physical characteristic of lithium-ion batteries.

Physical plausibility refers to how well generated observations follow known or predefined physical constraints and represent credible combinations of battery operating variables. In this study, PADS was specifically designed to promote physical plausibility by applying predefined constraints to variables such as charge rate, discharge rate, temperature, and state-of-charge conditions. However, physical plausibility itself was not independently measured as an outcome within TSDEF.

Feasibility refers to whether a synthetic data generation method is suitable for its intended research purpose. Since the evaluation used in this study primarily focuses on statistical fidelity, conclusions about feasibility are limited to synthetic dataset expansion under the statistical criteria examined by TSDEF. Therefore, the results do not establish whether either method is feasible for broader experimental, predictive, or safety-critical battery applications. These applications would require further testing and validation.

Realism is used as a broader description of data that demonstrates both appropriate statistical and physical characteristics. Since statistical fidelity and physical plausibility represent different dimensions of synthetic data quality, realism is avoided when only one of these dimensions is actually being evaluated. This distinction is important because a dataset can closely reproduce the statistical properties of the original data without necessarily representing all aspects of real-world battery behavior.

A synthetic data generation method refers to the overall approach used to create synthetic observations. Therefore, PADS and DDS are described throughout this study as synthetic data generation methods. Moreover, a generator refers more specifically to the implemented computational procedure that produces synthetic records according to one of these methods. For this reason, the implementations of PADS and DDS may be described as generators when discussing the actual process of producing synthetic data.

A model refers to a mathematical or statistical

representation of relationships between variables. Within this study, the Gaussian copula fitted as part of DDS and the ordinary least-squares regression specifications used in Level 3 are considered models. PADS and DDS themselves are not generally referred to as models when discussing the overall approaches used for synthetic data generation. Keeping this distinction helps separate the generation methods from the mathematical and statistical models used within them. Finally, an evaluation framework refers to an organized procedure that combines multiple evaluation levels and measurements to assess synthetic data quality. TSDEF is therefore referred to as the evaluation framework. Individual measurements within TSDEF, including mean deviation, correlation difference, Mean Absolute Correlation Error (MACE), and Mean Absolute Coefficient Difference (MACD), are instead referred to as evaluation metrics. Together, these definitions provide a consistent terminology for distinguishing how synthetic data are generated, how they are evaluated, and what conclusions can reasonably be drawn from the results.

Distributional Synthetic Data Generation

Within this study, two categories of synthetic generators will be reviewed: distributional approaches (statistical) and physics-aware approaches. To start, distributional methods attempt to replicate quantitative properties of existing datasets by modeling their underlying structure. According to (6), techniques that fit this category involve Gaussian-copula modeling and generative adversarial networks (GANs), which have all been widely used for synthesis. Unlike GAN-based generators, the DDS method evaluated in this study belongs to the Gaussian-copula family of distributional synthesis methods. Rather than learning data distributions through a neural network, DDS models the empirical statistical dependence structure of the original dataset using a Gaussian copula before generating synthetic samples. Similarly, (7) found that generative models are very efficient at reproducing complex patterns in data structures. However, while both studies reached the same conclusion, their findings cover a wide range of synthesis studies. LIB data is unique, as synthetic data generation methods have to not only reproduce complex data patterns but also maintain statistical fidelity. Therefore, since their studies were not LIB-specific, it is difficult to assert confidence in their findings.

Furthermore, other studies' results have challenged the findings of the previous research papers. For instance, (8) argues that synthetic datasets can match surface-level distributions, but fail to preserve structural relationships, especially when working with smaller datasets. Their belief that distributional models may produce data that looks realistic but doesn't actually maintain underlying data behaviors is mirrored by others, too. Using previously mentioned GAN models, (8) found that distributional generative models trained on small-sample-size datasets often struggle to keep nonlinear interactions, leading to distortion in generated data. This limitation is significant in LIB battery research, where relationships between variables must not only mirror dataset patterns, but also follow physical laws. Comparisons between past studies on distributional synthetic data generation methods make it evident that there are conflicting opinions on its viability to generate meaningful data.

Physics-Aware Synthetic Data Generation

Moving on to the second model, physics-aware methods place an emphasis on realistic data over data structure. One widely studied example of this approach is the physics-informed neural network (PINN). Unlike the distinct physics-aware method evaluated in this study, which is a rule-based synthetic data generator that applies predefined physical constraints during sampling, PINNs embed governing physical equations into the training process of neural networks. Raissi and others introduce an example of a method like this called physics-informed neural networks (PINNs), which integrate differential equations into machine learning models to enforce physical laws (9). His study, although surface-level synthetic data generation method analysis, provided the framework for future studies on the topic. Focusing specifically on battery research, Richardson and others found that these PINN models can accurately simulate battery degradation behavior and provide realistic data even without experimental datasets (10). These synthesized beliefs illustrate that a physics-aware generator can artificially create large datasets that adhere to the laws of physics.

Despite their advantages, however, they are not without limitations. The studies are limited in scope in relation to synthesis, as PINN models were accurate in a controlled laboratory context, not a dataset expansion setting. Additionally, (11) notes that these physically constrained models aren't always able to fully capture the variability present in real-world data due to their reliance on simplified assumptions. Others echo this sentiment; while hybrid physics-machine learning models improve interpretability, they still struggle to replicate complex

systems. Therefore, the utilization of physics-aware methods presents a tradeoff between physical plausibility and statistical fidelity, as models that prioritize strict rules may deviate from the original dataset's distributions.

Evaluation of Synthetic Data Quality

Additionally, past explorations of data synthesis are often limited in how synthetic data is evaluated. Studies often utilized single-level distributional metrics such as Kolmogorov-Smirnov tests to determine similarity between real and synthetic datasets. However, others argue that these evaluation methods aren't enough to accurately determine a method's feasibility, as they only analyze a single dimension of data quality. Plenty of studies back this claim up, as well. For instance, a 2025 study by Santangelo *et al*. used a multilevel evaluation system called "SynthRO" to assess synthetic health data, which performed reasonably well (12). Unfortunately, there are no established evaluation methods in the LIB field, so a method must be developed.

Research Gap, Research Question, and Hypothesis

Existing literature emphasizes a research gap in the viability of two distinct synthetic data generation methods for LIB dataset expansion. Seeing few existing studies directly comparing and analyzing distributional synthesis and physics-aware synthesis, this current study will attempt to bridge this gap by synthesizing synthetic datasets and evaluating their accuracy through tests. Ultimately, the study will answer the question: Which synthetic data generation method—physics-aware or distributional data—is more feasible for expanding LIB datasets while preserving statistical fidelity?

The initial hypothesis for this study is that the distributional synthesis method will be more statistically aligned with the original dataset than the physics-aware one, and therefore better suited to expanding LIB datasets for statistical fidelity.

METHODS AND MATERIALS

To answer the research question, two synthetic data generation strategies were developed and analyzed. The first method, labeled *Physics-Aware Data Synthesis (PADS)*, utilized relative physics rules and electrochemical properties to generate data that aligns with physically plausible operating conditions. This method created new data that is not only consistent with the laws of physics but is also within engineering safety limits. The second approach, called *Distributional Data Synthesis (DDS)*, approached this problem in a different manner by utilizing statistics rather than physics-based constraints. By modeling the joint distribution of categorical and numerical variables using techniques such as Gaussian copulas, the generated samples reflect empirical relationships found in the real data (13).

This study also included a process to analyze the effectiveness of each synthetic data generation method in order to determine whether or not the strategy is truly feasible for real-world research applications. By utilizing a custom-made three-stage comparative framework, this paper analyzed each synthesized dataset from multiple perspectives. First, Level 1 calculated base statistical values by calculating the deviation of the mean, standard deviation, minimum, and maximum values of each numerical variable relative to the base dataset. Next, Level 2 evaluated correlation structures, comparing Pearson correlation coefficients between all numerical variable pairs and calculating pairwise differences to determine whether multivariate relationships are consistent. Finally, Level 3 evaluated multivariable regression structure by fitting the same ordinary least-squares (OLS) regression model separately to the original dataset (DF1), the PADS dataset (DF2), and the DDS dataset (DF3). Discharge rate (C) was used as the response variable, while nominal capacity, temperature, maximum state of charge, minimum state of charge, charge rate, cathode chemistry, and form factor were included as predictors. Regression coefficients and model diagnostics were then compared to evaluate whether the synthetic datasets preserved the regression relationships observed in the original data. Together, these three evaluation levels form the *Three-Level Synthetic Data Evaluation Framework (TSDEF)*, a consistent and accurate way to check how similar a synthesized dataset is to the original.

For replicability purposes, Appendix A links to and overviews the synthetic generation codes for both PADS and DDS, while the evaluation code for TSDEF can be found in Appendix B. In each respective appendix, specific implementation instructions are detailed.

Data Description

This study extracted and modified a public dataset from BatteryArchive.org (2026), a government-affiliated database that contains laboratory-acquired testing data for LIBs (14). The original dataset from BatteryArchive is referred to as DF1 and serves as the control dataset for this study. It comes from the "Cycling Data" page and contains 137 lithium-ion battery cells compiled from multiple published studies and institutional

research projects, including sources like Sandia National Laboratories, HNEI, CALCE, Oxford, and UL-PUR. BatteryArchive was cited as the download source, while the data sources were cited to follow the site usage guidelines.

Two synthetic datasets were generated from DF1: DF2, created using the PADS framework, and DF3, generated using the DDS statistical synthesis framework. For both of the synthesis methods, the datasets (DF2, DF3) were expanded to 5,000 records from the initial 137 records in DF1.

This study began with the six base operating condition variables: nominal capacity (Ah), operating temperature (°C), maximum state of charge (Max SOC), minimum state of charge (Min SOC), charge rate (C-rate), and discharge rate (C-rate). These were the primary variables that both synthetic data generation methods will attempt to replicate. Then, through mathematical modeling and analysis, further deviations to more complex statistics were done for multiple levels of analysis. The BatteryArchive website specifically added the variables of nominal cell capacity, temperature, SOC bounds, and C-rates, emphasizing their importance to the study of LIBs. Definitions of each variable are listed in Table 1.

Synthetic Data Generation Methods

Method 1: Physics-Aware Data Synthesis (PADS)

PADS generated synthetic datasets by sampling operating data from DF1, the BatteryArchive control dataset. This approach then synthetically expanded DF1 to 5,000 records by enforcing real-world physical constraints that stem from battery test practice and known LIB qualities. The ultimate purpose of PADS was to ensure that generated data stayed consistent with real-life expectations of battery qualities, even if inconsistencies began to pop up in the statistics.

Categorical variables (e.g., cathode type, form factor) were randomly selected from the original dataset categories (Table 2), while numerical values (e.g., temperature, SOC limits) were randomly sampled from realistic data value sets in DF1. For each of the 5,000 synthesized data points, a charge and discharge rate was given based on cathode type (Table 3). These values would later be adjusted using factors that account for other numerical values. In addition, invalid SOC combinations were removed by enforcing $SOC_{min} < SOC_{max}$.

In battery engineering, data variable quantities are not arbitrary: the average values for each variable

Table 1. *Lithium-ion battery operating variables included in the original and synthetic datasets.*

No.	Variable	Description
1	Ah	Nominal cell capacity, amount of energy a battery can deliver under specific conditions.
2	Temperature (°C)	The operating temperature during data testing, recorded in degrees Celsius.
3	Max SOC	The highest state-of-charge level reached during cycling, displayed as a percentage of the battery's full capacity.
4	Min SOC	The lowest state-of-charge level reached during cycling, displayed as a percentage of the battery's full capacity.
5	Charge Rate (C)	The rate at which the battery is charged, expressed in C-rate units.
6	Discharge Rate (C)	The rate at which the battery is discharged, expressed in C-rate units.

Table 2. *Predefined categorical and numerical sampling values used by the Physics-Aware Data Synthesis (PADS) generator.*

No.	Variable	Description
1	Cathode Type	LFP, NMC, NCA, LCO, NMC-LCO
2	Form Factor	18650, pouch, prismatic
3	Nominal Capacity (Ah)	0.74, 1.1, 1.35, 2.8, 3.0, 3.2, 3.4
4	Temperature (°C)	15, 23, 25, 35, 40
5	Max SOC (%)	100, 96.5, 80, 60
6	Min SOC (%)	0, 2.5, 20, 40

Table 3. *Baseline charge and discharge rates assigned to each cathode chemistry by the PADS generator.*

No.	Variable	Charge Rate (C)	Discharge Rate (C)
1	LFP	0.6	1.5
2	NMC	0.8	2.0
3	NCA	0.9	2.2
4	LCO	0.7	1.8
5	NMC-LCO	0.8	2.0

fall within an expected range due to either real-life constraints or laboratory safety regulations. Current rates are no exception, as they are bounded by concepts like polarization, heat, and laboratory failure mechanisms that are more likely to occur under aggressive charging or at unfavorable temperatures and state-of-charge (SOC) ranges (15). For these reasons, PADS imposed strict bounds on charge and discharge rates before conducting any further rule adjustments. Specifically, the charge and discharge rates were limited to the intervals

$$C_{chg} \in [0.2, 3.0] \quad \text{(Eq. 1)}$$

$$C_{dis} \in [0.2, 5.0] \quad \text{(Eq. 2)}$$

where C_{chg} represents the charge rate (C) between 0.2 and 3.0, and C_{dis} represents the discharge rate (C) between 0.2 and 5.0. These ranges were labeled as "engineering-feasible envelopes" by BatteryArchive, as the website standardized all their testing data to fall within the same range.

In addition, temperature was also a utilized variable to further constrain the permissible current of the synthesized data. Although the relationship between temperature and rate is not monotonic across all lab conditions (higher temperatures reduce resistance but also accelerate degradation), an increase in temperature typically reduces immediate rate limitations within a controlled testing band (16). Furthermore, low temperatures elevate plating risk during the charging process (17). With these facts in consideration, PADS models the maximum allowable charge rate as a linear graph around a reference temperature. The formula is modeled as

$$C_{chg,max}(T) = C_0 \cdot m(T),\ m(T) = (1 + 0.012(T - 25))$$
$$m(T) \text{ is clipped to } [0.8,1.3] \quad \text{(Eq. 3)}$$

where C_0 is a baseline rate deemed consistent with any cell operation simulations and the coefficient of 0.012 was chosen to limit LIB headroom temperature. This equation was derived by analyzing established relationships between temperature and charge rate and incorporating desired qualities into the formula, such as a linear slope.

The final bounds added for PADS are the SOC constraints. Oftentimes, real-world LIB charging begins to transition to a constant-voltage-limited area, a charge rate range where current begins to decrease—as SOC begins to approach the upper limit (An *et al*., 2019). To account for this factor in terms of the DF1 data trends, the limit is expressed as

$$C_{chg,eff} = C_{chg} \cdot g(SOC) \quad \text{(Eq. 4)}$$

where

$$g(SOC) = (1.2 - 0.003(SOC_{max} - SOC_{min}) \quad \text{(Eq. 5)}$$

In this equation, $g(SOC)$ is a monotonically decreasing function as SOC approaches SOC_{max} and the effective charge decreases. Theoretically, this constraint prevents PADS from producing unjustifiable samples with both high SOC rates and high charge rates (18).

Ultimately, PADS emphasizes plausibility above all else, accepting that marginal deviations may drift from DF1's frequencies. By prioritizing safe, reliable scenarios rather than statistical replicas of the control dataset, PADS explores the possibility that synthesized data can be both statistically accurate and physically plausible.

Method 2: Distributional Data Synthesis (DDS)

DDS approached this problem in the completely opposite way by prioritizing replicating the statistical values pulled from DF1. With the goal of preserving the joint structure of DF1 as closely as possible, DDS synthetically expanded DF1's dataset of 137 cell records into 5,000 records. DDS was produced based on the belief that although Battery Archive had a relatively small archive of LIB testing data, it still contains valuable information in how variables interact with one another. This sentiment is echoed by the general research community, as open synthetic data research has repeatedly found that explicitly modeling multivariate dependence often outperforms individual sampling because it maintains relationships that downstream models are reliant on (19).

5,000 DDS samples were generated computationally by estimating the distribution in DF1 using the Gaussian copula method. In order to simplify the study of multiple base variables (Table 1), DDS combines all the numerical variables in the dataset into a vector. Let $X = (X_1,...,X_p)$ denote the vector of all these base variables (capacity, temperature, SOC bounds, C-rates). Through the utilization of a Gaussian copula, each marginal variable is turned into a uniform variable (n = # of observations) using an empirical rank transformation:

$$U_i = \frac{rank(X_i)}{n+1} \quad \text{(Eq. 6)}$$

Next, they were put in a latent normal space through the inverse Gaussian transformation:

$$Z_i = \sqrt{2} erfinv(2U_i - 1) \quad \text{(Eq. 7)}$$

Afterward, an estimated correlation matrix captured the dependence of that variable:

$$Z_{syn} \sim N(\mu, \sum),\ \mu = mean\ vector,\ \sum = variance\ matrix \quad \text{(Eq. 8)}$$

To put it simply, this process makes it easier to analyze each individual variable by separating it from its relationships with other variables. This approach was done because it is a prevalent strategy that labels Gaussian copulas as a strong foundation for data generation. After taking samples from each generated value in the latent space, they are reverted back to their original feature scales through inverse transformation. This is imperative because while the transformation to matrix form was necessary for DDS, the actual DF3 dataset must be in the same form as DF1 to sustain accurate comparisons (20).

Ultimately, DDS's top priority was to replicate the statistical values pulled from DF1. By emphasizing the importance of data similarity and reliability, DDS provides the potential to synthetically expand existing datasets with extreme accuracy.

Three-Level Synthetic Data Evaluation Framework (TSDEF)

Level 1: Basic Statistical Distribution Evaluation

At the base level, each quantitative variable was analyzed using standard descriptive statistics and compared with the original dataset (21). Assume each variable is represented by variable X with n observations. Let

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{(Eq. 9)}$$

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (X_i - \mu_X)^2} \quad \text{(Eq. 10)}$$

where μ_X symbolizes the mean and σ_X provides the sample standard deviation. In addition, the minimum and maximum values for each numerical variable were also recorded in order to prevent any of the synthesis methods from producing impossible values. To simplify Level 1 for analysis, let S = the original variables. The variables were then compared to the original dataset, and a final deviation will be calculated through the expression

$$\Delta S = S_{syn} - S_{orig} \quad \text{(Eq. 11)}$$

This first level of analysis provided statistics on whether the individual synthetic variables seem realistic from a surface-level perspective. However, this study recognizes that univariate similarity alone can not guarantee full relational accuracy between the synthetic values, which is why further levels of evaluation are required (21).

Level 2: Correlation Structure Preservation

This second evaluation level delves deeper into statistics by analyzing multivariate dependencies among the base variables. Level 2 makes use of the Pearson correlation coefficient, a commonly applied method of analysis for scientific research, to measure linear dependence between two variables (21). For variables X and Y, Pearson's correlation coefficient is calculated as

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(X_i - \mu_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i - \mu_Y)^2}} \quad \text{(Eq. 12)}$$

The Pearson coefficient ranges from -1 to 1, with -1 representing an inverse relationship and 1 representing a linear relationship. This equation utilizes the mean of each variable (pulled from Level 1) as well as the spread (variability) of each variable.

To determine the Pearson coefficient's deviation between the synthetic and original dataset, the pairwise correlation difference was used, defined as

$$\Delta\rho_{ij} = \rho_{ij}^{syn} - \rho_{ij}^{orig} \quad \text{(Eq. 13)}$$

The format mirrors that of Level 1's deviation metric. This study utilizes it because it is simple to use and is widely regarded as one of the best matrix norms in data research (22). To complement the pairwise correlation differences, an overall correlation-matrix error was also calculated using the Mean Absolute Correlation Error (MACE). This metric summarizes the average absolute deviation between the original and synthetic Pearson correlation coefficients across all unique variable pairs:

$$MACE = \frac{1}{N}\sum_{i<j} \left| \rho^{syn}{}_{ij} - \rho^{orig}{}_{ij} \right| \quad \text{(Eq.14)}$$

Where N is the total number of variable pairs, $\rho^{syn}{}_{ij}$ is the Pearson correlation coefficient of the synthetic dataset, and $\rho^{orig}{}_{ij}$ is the Pearson correlation coefficient of the original dataset. A lower MACE value represents a better preservation of the correlation structure.

Level 3: Regression Analysis

Level 3 evaluated whether the multivariable statistical relationships observed in the original dataset were similarly represented in the two synthetic datasets using ordinary least squares (OLS) regression. The same regression specification was fitted separately to DF1, DF2, and DF3 to ensure that coefficient estimates were directly comparable across datasets. The response variable was discharge rate (C). Five numerical predictors were included: nominal capacity (Ah), operating temperature (°C), maximum state of charge (Max SOC), minimum state of charge (Min SOC), and charge rate (C). Cathode chemistry and form factor were included as categorical predictors.

Categorical predictors were represented using treatment, or dummy, coding. LCO was specified as the reference category for cathode chemistry, while 18650 was specified as the reference category for form factor. Therefore, four cathode indicator variables represented LFP, NCA, NMC, and NMC-LCO relative to LCO, while two form-factor indicator variables represented pouch and prismatic cells relative to 18650 cells. An intercept was included in every fitted model.

For observation i, the regression model was

$$\begin{aligned} DischargeRate_i = {} & \beta_0 + \beta_1(Ah_i) + \beta_2(Temperature_i) \\ & + \beta_3(MaxSOC_i) + \beta_4(MinSOC_i) + \beta_5(ChargeRate_i) \\ & + \beta_6 I(Cathode_i = LFP) + \beta_7 I(Cathode_i = NCA) \\ & + \beta_8 I(Cathode_i = NMC) + \beta_9 I(Cathode_i = NMC \\ & - LCO) + \beta_{10} I(FormFactor_i = pouch) \\ & + \beta_{11} I(FormFactor_i = prismatic) + \varepsilon_i \end{aligned} \quad \text{(Eq. 15)}$$

where β_0 is the intercept, β_1 through β_5 are coefficients for the predictors, and $I(x)$ represents an indicator variable of 1 when the condition is satisfied and 0 when it is not. ε_i is the residual error. Finally, since LCO and 18650 were utilized as reference categories, they were used without separate indicator coefficients.

The coefficients of the numerical predictors represent the estimated change in discharge rate when one unit variable is changed, and all other variables are held constant. Coefficients for categorical predictors represent the estimated difference in discharge rate relative to the category. The study kept the intercept as part of the mathematical equation but was not given a physical interpretation because the chance that simultaneous zero values occur over all numerical predictors was almost impossible in LIB operating conditions.

However, before regression fitting, all numerical variables were converted to numeric format, with non-convertible entries deleted. Observations missing the response variable or any predictor were removed using complete-case analysis. The resulting regression sample sizes for DF1, DF2, and DF3 were all recorded separately and can be found in Appendix D3.

The regression reporting included coefficient estimates, standard errors, and 95% confidence intervals for all terms, including the intercept. The overall model fit was summarized using R^2, adjusted R^2, and residual standard error. All regression assumptions were evaluated using normal Q-Q plots, the Breusch-Pagan/Koenker test, and Cook's distance for influential observations. This study assesses multicollinearity using variance inflation factors (VIFs) for each non-intercept design-matrix column, as well as the rank and condition number of the regression design matrix.

The formula for VIF is

$$VIF_j = \frac{1}{1-R^2_j} \quad \text{(Eq. 16)}$$

where R^2_j is calculated by regressing variable j on the other predictors. A larger VIF value indicates that the variance of a numerical predictor's regression value may have been inflated by the other variables.

Since the objective of this level was to compare fitted statistical relationships rather than predictive performance, the regression results were interpreted as a measure of structural preservation rather than as evidence of predictive accuracy.

Finally, to evaluate how synthetic data actually affected these relationships, the study once again used differences to determine the coefficient deviation of the synthetic dataset.

$$\Delta\beta_k = \beta_k^{syn} - \beta_k^{orig} \quad \text{(Eq. 17)}$$

This metric measures the difference between the estimated coefficient from the synthetic dataset and the corresponding coefficient from the real control dataset. By focusing on the overall structure of the relationships between variables, regression modeling can accurately determine whether or not a synthetic dataset is realistic.

To summarize the overall similarity between regression models, the Mean Absolute Coefficient Difference (MACD) was calculated as the average absolute difference between corresponding regression coefficients in the original and synthetic datasets. Unlike signed averages, this measure prevents positive and negative coefficient differences from canceling each other and therefore provides a more representative measure

of overall coefficient deviation. Lower MACD values indicate greater similarity between the fitted regression models.

RESULTS

Results Overview

As mentioned earlier in the methodology section, this study expanded the original dataset of 137 cells into synthetic datasets with 5,000 cells. Therefore, the quantity of data utilized is far too large to display within this section. All raw data table sheets for each of the synthetic methods can be found in Appendix C, while the raw evaluation data tables as well as Level 3 graphs can be found in Appendix D.

In this section, the study summarized the findings from comparing the original dataset DF1 with the synthesized datasets DF2 and DF3. To conduct the evaluations, the quantitative results presented by TSDEF were used. Thus, this section was separated into three distinct sections, each representing a different level of evaluation.

Level 1 | Basic Evaluation Results

Based on the Level 1 results (Table 4), it was clear that DF3 is the closest match to DF1 for all six of the operating-condition variables that were analyzed. The evaluated mean, standard deviations, and ranges were almost identical between DF1 and DF3, which reflected DF3's emphasis on quantitative statistical fidelity. A clear example of this is shown through the discharge rate: DF1 reported a mean of 1.27 C, DF2 reported a mean of 1.93 C, and DF3 reported a mean of 1.26 C. This is mirrored in the deviation calculation ($\Delta S \approx -0.01$), while DF2 showed a larger deviation ($\Delta S \approx +0.66$). This is seen throughout the mean and standard deviation values for the operating variables. DF2 also had a lower mean Max SOC and a higher mean Min SOC than DF1, meaning the allowed SOC values are narrowed. Despite this, DF2 shared the same values as DF1 and DF3 across many ranges, thereby providing it with greater synthetic relevance.

Level 2 | Advanced Multivariate Dependency Results

The multivariate dependency data from Level 2's results (Table 5) clearly depicted how well DF2 and DF3 preserved important relationships between variables (Variable 1 and Variable 2) through the all-in-one metric, "correlation difference ($\Delta\rho$)". A larger $\Delta\rho$ value often indicated that original relationships are not preserved. Across the DF1 vs. DF2 comparisons, DF2 showed significantly larger Pearson correlation deviations than the original dataset. The most extreme case, the relationship between Max SOC and Min SOC, had a

Table 4. *Comparison of descriptive statistics for the six operating variables in the original, PADS-generated, and DDS-generated datasets.*

Variable	Ah	Temperature (°C)	Max SOC	Min SOC	Charge Rate (C)	Discharge Rate (C)
DF1: mean	2.44	26.01	94.61	5.31	0.59	1.27
DF2: mean	2.25	27.44	84.41	15.14	0.77	1.93
DF3: mean	2.44	25.99	94.75	4.93	0.57	1.26
DF1: std	0.98	5.91	11.55	11.57	0.35	0.82
DF2: std	1.03	8.82	15.87	15.94	0.18	0.40
DF3: std	0.97	5.68	11.40	11.02	0.32	0.80
DF1: min	0.74	15.00	60.00	0.00	0.50	0.50
DF2: min	0.74	15.00	60.00	0.00	0.34	0.50
DF3: min	0.74	15.00	60.00	0.00	0.50	0.50
DF1: max	3.40	40.00	100.00	40.00	2.00	3.00
DF2: max	3.40	40.00	100.00	40.00	2.00	3.21
DF3: max	3.40	40.00	100.00	40.00	2.00	3.00

Notes: DF1 = original dataset; DF2 = physics-aware data synthetic dataset; DF3 = distributional data synthetic dataset; std = standard deviation.

$\Delta\rho$ of +0.97, showing a large change from the original inverse relationship. The more-often-than-not deviations indicate that many of the original correlations are not preserved in DF2.

In contrast, DF3 produced smaller deviations than DF2, indicating that the overall multivariate relationship was maintained better. Several relationships, such as Ah vs temperature ($\Delta\rho$ = +0.07) and discharge rate vs

Table 5. *Pairwise Pearson correlations and correlation differences between the original and synthetic datasets.*

Original	Compare	Variable 1	Variable 2	(Synthetic Correlation)	(Original Correlation)	(Correlation Difference)
DF1	DF2	Ah	Temperature (°C)	0.01	-0.31	0.32
DF1	DF2	Ah	Max SOC	-0.01	0.02	-0.03
DF1	DF2	Ah	Min SOC	-0.01	-0.03	0.02
DF1	DF2	Ah	Charge Rate (C)	-0.02	-0.43	0.42
DF1	DF2	Ah	Discharge Rate (C)	-0.02	-0.24	0.21
DF1	DF2	Max SOC	Temperature (°C)	-0.01	0.09	-0.10
DF1	DF2	Max SOC	Min SOC	-0.03	-1.00	0.97
DF1	DF2	Min SOC	Temperature (°C)	0.01	-0.08	0.10
DF1	DF2	Charge Rate (C)	Temperature (°C)	0.50	0.59	-0.09
DF1	DF2	Charge Rate (C)	Max SOC	-0.23	0.12	-0.35
DF1	DF2	Charge Rate (C)	Min SOC	0.21	-0.11	0.33
DF1	DF2	Charge Rate (C)	Discharge Rate (C)	0.76	0.18	0.58
DF1	DF2	Discharge Rate (C)	Temperature (°C)	0.52	0.13	0.39
DF1	DF2	Discharge Rate (C)	Max SOC	-0.26	0.01	-0.27
DF1	DF2	Discharge Rate (C)	Min SOC	0.25	-0.01	0.26
DF1	DF3	Ah	Temperature (°C)	-0.24	-0.31	0.07
DF1	DF3	Ah	Max SOC	0.13	0.02	0.11
DF1	DF3	Ah	Min SOC	0.19	-0.03	0.22
DF1	DF3	Ah	Charge Rate (C)	0.10	-0.43	0.53
DF1	DF3	Ah	Discharge Rate (C)	-0.19	-0.24	0.05
DF1	DF3	Max SOC	Temperature (°C)	0.02	0.09	-0.07
DF1	DF3	Max SOC	Min SOC	-0.31	-1.00	0.69
DF1	DF3	Min SOC	Temperature (°C)	-0.09	-0.08	0.00
DF1	DF3	Charge Rate (C)	Temperature (°C)	0.10	0.59	-0.50
DF1	DF3	Charge Rate (C)	Max SOC	-0.01	0.12	-0.13
DF1	DF3	Charge Rate (C)	Min SOC	0.25	-0.11	0.36
DF1	DF3	Charge Rate (C)	Discharge Rate (C)	0.07	0.18	-0.10
DF1	DF3	Discharge Rate (C)	Temperature (°C)	0.16	0.13	0.03
DF1	DF3	Discharge Rate (C)	Max SOC	0.07	0.01	0.05
DF1	DF3	Discharge Rate (C)	Min SOC	0.00	-0.01	0.00

Notes: DF1 = original dataset; DF2 = physics-aware synthetic dataset; DF3 = distributional data synthetic dataset.

temperature ($\Delta\rho$ = +0.03), remained closely aligned with the original dataset. However, several moderate-to-large deviations were also observed, including Ah vs. Charge Rate (Δρ = +0.53), Charge Rate vs. Temperature (Δρ = −0.50), Charge Rate vs. Min SOC (Δρ = +0.36), and Max SOC vs. Min SOC (Δρ = +0.69). Therefore, while DF3 preserved the correlation structure more effectively than DF2, the results indicate improved rather than complete preservation of multivariate relationships. The Mean Absolute Correlation Error likewise showed lower overall correlation error for DF3 than for DF2 (Table 6).

Table 6. *Mean Absolute Correlation Error for the PADS- and DDS-generated datasets relative to the original dataset.*

Dataset	MACE
DF1 vs. DF2	0.296
DF1 vs. DF3	0.195

Across all 15 unique variable pairs, DF2 produced a Mean Absolute Correlation Error (MACE) of 0.296, whereas DF3 produced a lower MACE of 0.195. Thus, DF3 reduced the overall pairwise correlation error by approximately 34.1% relative to DF2. Although this result indicated that DF3 preserved the source correlation structure more effectively overall, the remaining MACE and several large pairwise deviations demonstrate that preservation was incomplete.

Level 3 | Regression Analysis Results

Level 3 compared the same OLS regression model across the original dataset (DF1), the PADS synthetic dataset (DF2), and the DDS synthetic dataset (DF3). The regression included 137 observations for DF1 and 5,000 observations each for DF2 and DF3. Complete coefficient estimates, standard errors, and 95% confidence intervals can be found in Table 7. Overall, the model fit differed greatly across the three datasets. DF1 produced an of 0.299 and an adjusted R^2 of 0.249, while DF3 had an R^2 of 0.106 and an adjusted R^2 of 0.105. In contrast, DF2 produced an extremely high R^2 of 0.999918 and an adjusted R^2 of 0.999918. At first, this near-perfect fit may seem like PADS performed much better. However, it should not be interpreted as independent evidence of better predictive performance. Since PADS generates operating variables through specific rule-based relationships, the extremely high R^2 most likely come from the structured relationships that were already introduced during the synthesis process.

Table 7. *Ordinary least-squares regression coefficient estimates for discharge rate in the original, PADS-generated, and DDS-generated datasets.*

Dataset	Predictor	β	SE	95% CI
DF1	Intercept	8.493	24.283	[-39.559, 56.546]
DF1	Cathode: LFP vs. LCO	-1.351	4.771	[-10.791, 8.090]
DF1	Cathode: NCA vs. LCO	5.081	6.634	[-8.047, 18.209]
DF1	Cathode: NMC vs. LCO	4.915	6.405	[-7.760, 17.590]
DF1	Cathode: NMC-LCO vs. LCO	4.125	6.200	[-8.145, 16.394]
DF1	Form Factor: pouch vs. 18650	-2.629	4.302	[-11.143, 5.884]
DF1	Form Factor: prismatic vs. 18650	-1.647	4.942	[-11.428, 8.133]
DF1	Ah	-3.291	1.356	[-5.975, -0.607]
DF1	Temperature (°C)	-0.009	0.013	[-0.035, 0.016]
DF1	Max SOC	-0.018	0.309	[-0.629, 0.593]
DF1	Min SOC	-0.026	0.309	[-0.637, 0.586]
DF1	Charge Rate (C)	0.302	5.767	[-11.108, 11.713]
DF2	Intercept	0.166	0.002	[0.161, 0.170]
DF2	Cathode: LFP vs. LCO	-0.069	<0.001	[-0.069, -0.068]

Continued Table 7. *Ordinary least-squares regression coefficient estimates for discharge rate in the original, PADS-generated, and DDS-generated datasets.*

Dataset	Predictor	β	SE	95% CI
DF2	Cathode: NCA vs. LCO	-0.067	0.001	[-0.068, -0.066]
DF2	Cathode: NMC vs. LCO	-0.034	<0.001	[-0.034, -0.033]
DF2	Cathode: NMC-LCO vs. LCO	-0.034	<0.001	[-0.034, -0.033]
DF2	Form Factor: pouch vs. 18650	0.000	<0.001	[-0.000, 0.000]
DF2	Form Factor: prismatic vs. 18650	0.000	<0.001	[-0.000, 0.000]
DF2	Ah	-0.000	<0.001	[-0.000, 0.000]
DF2	Temperature (°C)	0.001	<0.001	[0.001, 0.001]
DF2	Max SOC	-0.000	<0.001	[-0.000, -0.000]
DF2	Min SOC	0.000	<0.001	[0.000, 0.000]
DF2	Charge Rate (C)	2.328	0.003	[2.321, 2.334]
DF3	Intercept	0.587	0.107	[0.376, 0.797]
DF3	Cathode: LFP vs. LCO	0.270	0.042	[0.189, 0.352]
DF3	Cathode: NCA vs. LCO	0.448	0.041	[0.367, 0.529]
DF3	Cathode: NMC vs. LCO	0.661	0.048	[0.566, 0.755]
DF3	Cathode: NMC-LCO vs. LCO	0.114	0.047	[0.022, 0.206]
DF3	Form Factor: pouch vs. 18650	-0.083	0.031	[-0.143, -0.023]
DF3	Form Factor: prismatic vs. 18650	-0.152	0.053	[-0.256, -0.049]
DF3	Ah	-0.244	0.013	[-0.270, -0.219]
DF3	Temperature (°C)	0.011	0.002	[0.007, 0.015]
DF3	Max SOC	0.006	0.001	[0.004, 0.008]
DF3	Min SOC	-0.002	0.001	[-0.005, -0.000]
DF3	Charge Rate (C)	0.169	0.030	[0.110, 0.227]

Notes: DF1 = original dataset; DF2 = PADS synthetic dataset; DF3 = DDS synthetic dataset. LCO was the reference category for cathode chemistry, and 18650 was the reference category for form factor. Standard errors and 95% confidence intervals are conventional OLS estimates.

Table 8. *Sample sizes and ordinary least-squares model-fit statistics for the original, PADS-generated, and DDS-generated datasets.*

Dataset	*n*	R^2	Adjusted R^2	Residual Standard Error
DF1	137	0.299	0.249	0.706
DF2	5,000	1.000	1.000	0.003
DF3	5,000	0.106	0.105	0.752

Table 9. *Diagnostic statistics for the ordinary least-squares regression models fitted to the original, PADS-generated, and DDS-generated datasets.*

Dataset	BP p-value	JB p-value	Maximum Cook's D	Maximum VIF	Rank	Condition Number
DF1	< 0.001	< 0.001	0.0534	inf	10/12	8.32×10^{17}
DF2	< 0.001	< 0.001	0.0055	118.87	12/12	9.00×10^{3}
DF3	< 0.001	< 0.001	0.0045	inf	11/12	3.95×10^{16}

Notes: BP = Breusch-Pagan heteroscedasticity test; JB = Jarque-Bera residual-normality test; VIF = variance inflation factor. Infinite VIF values indicate exact linear dependence among predictors. Rank lower than the total number of design-matrix columns indicates rank deficiency.

Table 10. *Differences between regression coefficients estimated from each synthetic dataset and those estimated from the original dataset.*

Predictor	DF2 Δβ	DF3 Δβ
Cathode: LFP	+1.282	+1.621
Cathode: NCA	-5.148	-4.633
Cathode: NMC	-4.949	-4.254
Cathode: NMC-LCO	-4.158	-4.011
Form Factor: pouch	+2.629	+2.546
Form Factor: prismatic	+1.647	+1.495
Ah	+3.291	+3.047
Temperature	+0.011	+0.020
Max SOC	+0.018	+0.024
Min SOC	+0.026	+0.023
Charge Rate	+2.025	-0.134

Table 11. *Mean Absolute Coefficient Difference for the PADS- and DDS-generated datasets relative to the original dataset.*

Dataset	MACD
DF1 vs DF2	2.289
DF1 vs DF3	1.983

Notes: DF1 = original dataset; DF2 = physics-aware data synthetic dataset; DF3 = distributional data synthetic dataset.

The coefficient estimates also showed major differences between the datasets. In DF1, several coefficients had large standard errors and wide confidence intervals. For example, the NCA cathode coefficient was 5.081 with a standard error of 6.634 and a 95% confidence interval from -8.047 to 18.209. The DF1 Ah coefficient was -3.291, with a standard error of 1.356 and a 95% confidence interval from -5.975 to -0.607. On the other hand, many of the DF2 coefficients had extremely small standard errors. This is consistent with the highly constrained structure of the PADS dataset. DF3 also had noticeable differences from the original coefficients, although overall, its coefficients were somewhat closer to DF1 than those from DF2.

To summarize these differences, the Mean Absolute Coefficient Difference (MACD) was used to measure overall coefficient preservation, excluding the intercept. DF2 produced a MACD of 2.289, while DF3 had a lower MACD of 1.983. Based on this measurement, DF3 preserved DF1's regression coefficients slightly better than DF2. However, this result cannot be looked at on its own. Further diagnostics found severe multicollinearity and rank deficiency in the DF1 and DF3 models, which makes some of the individual coefficients unstable. The largest differences were mostly seen in cathode type, form factor, and nominal capacity (Ah). Meanwhile, temperature and the SOC variables generally had much smaller differences.

The residual diagnostics also revealed several problems with the fitted regression models. Breusch-Pagan tests were significant for DF1, DF2, and DF3 ($p < 0.001$ for each), showing evidence of heteroscedasticity in all three datasets. Jarque-Bera tests were also significant for all three models ($p < 0.001$), indicating that the residuals did not fully follow a normal distribution. The graphs showed similar issues. The residual-versus-fitted plots had visible patterns instead of completely random residuals. DF2 had especially organized residual patterns despite having a near-perfect R-squared, while DF3 showed noticeable parallel banding. In addition, the DF3 normal Q-Q plot curved away from the theoretical reference line. Together, these results show that the OLS assumptions were not fully satisfied. Maximum Cook's-

distance values were 0.0534 for DF1, 0.0055 for DF2, and 0.0045 for DF3. However, these values were not used alone to judge the models and were instead considered along with the other regression diagnostics.

The multicollinearity results revealed an even more serious issue with the regression specification. DF1 had a design-matrix rank of 10 out of 12 columns and a condition number of 8.32×10^{17}, showing both rank deficiency and severe ill-conditioning. Multiple categorical variables and Charge Rate had infinite VIF values. Max SOC and Min SOC were also extremely high, with VIF values of 3472.76 and 3489.42. DF2 was different because its model was full rank at 12/12, but it still had substantial multicollinearity. For example, Charge Rate had a VIF of 118.87, while Temperature had a VIF of 30.28. DF3 also had problems, with a rank of 11 out of 12 columns and a condition number of 3.95×10^{16}. Infinite VIF values were found for Charge Rate and the pouch form-factor variable. These results make it clear that several individual regression coefficients are not completely stable. Because of this, the coefficients are better used as exploratory comparisons between the datasets rather than as precise measurements of the independent effect of each predictor.

Taken together, Level 3 showed much weaker preservation than Levels 1 and 2. DF3 did produce a lower MACD than DF2, meaning that its coefficients were overall closer to those of DF1 under this specific regression model. However, neither synthetic dataset was able to reproduce a regression structure that could be considered fully stable or compliant with the OLS assumptions. This is an important limitation because it shows that similarity at the earlier evaluation levels does not necessarily mean that more complex statistical relationships will also be preserved. Therefore, the Level 3 results are best viewed as an exploratory comparison of regression structure rather than evidence that either synthetic dataset has better predictive performance.

DISCUSSION

The results of TSDEF reveal a clear tradeoff between generating data through statistical fidelity and physics-based constraints, as shown when comparing PADS and DDS's synthetic datasets. Although both methods successfully expand DF1 into 5,000 artificial records, their actual feasibility remains largely dependent on which aspect of data fidelity is prioritized.

At Level 1, DDS is consistent with the original dataset in terms of surface-level dataset attributes like mean, standard deviation, and range (Table 4). As such, the minimal deviations reinforce DDS's prioritization of maintaining the original dataset's empirical structure. Because DDS is instructed to explicitly base the empirical distribution on DF1, it was expected that these base statistics would remain nearly identical. These results strengthen past research findings, highlighting how distribution-based generators are an effective way to replicate data characteristics (23). Next, Level 2's analysis on the relationships between variables in DF3 is mostly similar to the findings of Level 1: many of the correlations were preserved, often with little to no deviation (Table 4). However, some exceptions, like the relationship between charge rate and temperature ($\Delta\rho$ = -0.50), reveal how certain dependencies aren't perfectly maintained. Indeed, this limitation illuminates a known challenge in copula-based methods where nonlinear relationships aren't always kept (19). Despite this, DDS's Level 1 and Level 2 were still overall a good representation of the original dataset's qualities. Although DDS generally preserved pairwise correlations more effectively than PADS, several relationships still exhibited substantial deviations. In particular, the Ah–Charge Rate ($\Delta\rho = +0.53$), Charge Rate–Temperature ($\Delta\rho = -0.50$), Charge Rate–Min SOC ($\Delta\rho = +0.36$), and Max SOC–Min SOC ($\Delta\rho = +0.69$) relationships indicate that not all multivariate dependencies were maintained. Consequently, the Level 2 results should be interpreted as evidence of improved correlation preservation relative to PADS rather than complete preservation of the original multivariate structure.

One notable finding was the perfect inverse correlation ($\rho = -1.00$) between maximum SOC and minimum SOC in the original BatteryArchive dataset. This relationship is likely to reflect the construction of the source dataset rather than a universal electrochemical property of lithium-ion batteries. Because BatteryArchive contains predefined operating SOC windows, maximum and minimum SOC values occur in deterministic pairs. Therefore, deviations involving this variable pair should be interpreted cautiously and should not be generalized as evidence of broader battery behavior.

In contrast, PADS produced consistently larger deviations across the first two evaluation levels, sometimes almost to the point where empirical and dependent structures completely cease to exist ($\Delta\rho \approx \pm 1.00$). However, it's important to note that these deviations aren't necessarily evidence of poor performance but rather a reflection of the fundamental design of PADS. Because PADS prioritizes physics-

based constraints over statistical fidelity, certain structures and relationships are forcibly abandoned in favor of plausibility (Table 5). The easiest example is the multivariate dependency deviation between Min SOC and Max SOC ($\Delta\rho = 0.97$). By narrowing the SOC range, far beyond DF1's range, DF2 better aligns itself with electrochemical behaviors and safety constraints. This can be applied to most noticeable deviations between DF1 and DF3 (15).

Level 3 provided a more complicated picture than Levels 1 and 2. Under the fitted OLS model, DDS had a lower overall coefficient deviation from DF1 than PADS, with MACD values of 1.983 and 2.289, respectively. However, the diagnostic results showed that these coefficient differences could not be looked at separately from the instability of the models. Both DF1 and DF3 were rank-deficient, and several predictors had extremely large or even infinite VIF values. One of the clearest examples is the near-perfect inverse relationship between Max SOC and Min SOC observed in Level 2. In DF1, this relationship resulted in VIF values above 3,400 for both variables. Because of this, the individual coefficients in the original model are sensitive to predictor redundancy and should not be treated as completely stable measurements of independent battery relationships.

PADS had a different issue. Unlike DF1 and DF3, its model was full rank, but it produced an R^2 of approximately 0.9999 along with extremely small residual errors. Although this initially appears to be a strong result, it does not necessarily mean that PADS provides better predictive data. PADS creates variables using specific rule-based relationships, so the regression partially finds relationships already built into the synthetic generation process. This can also be seen in the structured residual patterns of DF2 and its remaining high VIF values. Therefore, an extremely high R^2 does not automatically mean that the original multivariable statistical structure was realistically preserved.

Overall, Level 3 further shows that the quality of synthetic data depends heavily on which part of the data is being evaluated. DDS performed better in Levels 1 and 2 and also produced the smaller MACD in Level 3. However, neither synthetic dataset was able to fully preserve a stable regression structure. Therefore, the Level 3 results are better interpreted as showing limitations in structural preservation rather than ranking either method based on predictive performance.

Ultimately, however, the initial hypothesis was correct, as the statistics present DDS as the overall superior synthetic data generation method for this study. The findings from this study mirror prior studies on the two synthetic generation methods, as DDS effectively preserved DF1's statistical structure across TSDEF's first two evaluation tests.

CONCLUSION

These findings have multiple implications for the future of synthetic generation in LIB research. First, the study's results showcase how synthetic data generation can help address the prominent bottleneck of limited data in the LIB research field. Since real-world lab battery tests are expensive and time-consuming, this newfound ability to expand DF1 from 137 cells to 5,000 suggests that large-scale synthetic data analysis is possible with further research on this topic. If the method proves to be reliable following further research, this LIB data increase will allow researchers to use machine learning to advance LIB knowledge. This would be significant. Furthermore, the results from this study reinforce the idea that the usefulness of a synthesis method relies on the purpose of the initial dataset. For example, DDS thrives when aiming to mimic statistical structure, while PADS is successful in maintaining realistic conditions for operating variables. Thus, while DDS is the better option for this study, researchers should not assume that one method is always better for any given scenario. Finally, this study asserts that when analyzing the reliability of synthesis data, one should go beyond surface-level similarities. Although a synthetic dataset may seem realistic at first, deeper observations may reveal structural distortions, as shown through TSDEF.

However, this study also has some limitations that may skew interpreted results in nuanced ways. The first major limitation was the fact that the original dataset pulled from BatteryArchive is relatively small, as it only contains 137 data cells. Even though this problem is the exact reason why synthetic expansion is needed, it still limits the information that can be gathered from statistical analysis. After all, the methods of synthetic generation are based on data samples from DF1 and, therefore, are generated from just 137 samples. Because DF1 doesn't fully represent the entirety of LIB behavior, DF2 and DF3 may reproduce only a portion of real-world data structures.

Additionally, the study only investigates six base operating variables. Although each of them possesses meaningful and relevant quantitative data, they don't fully capture the complexities of LIBs. Other variables,

such as internal resistance or battery chemistry, were not included in order to narrow the focus of this study. By doing this, the study sacrifices potentially data-changing statistics in favor of simplicity and regulation, which can limit the study's understanding of LIBs.

A major limitation of Level 3 is the instability of the selected OLS model. The formal diagnostics found rank deficiency in both DF1 and DF3, along with severe multicollinearity among several predictors. Again, the strongest example was the relationship between Max SOC and Min SOC in DF1, where both variables had VIF values above 3,400. This means that individual coefficients from the affected models cannot be cleanly separated and interpreted as independent predictor effects. Instead, Level 3 is more useful as an exploratory comparison of how the overall fitted coefficient structure changes after synthetic generation.

The residual diagnostics also found evidence of heteroscedasticity and departures from normality in all three models. Because of this, the conventional OLS uncertainty estimates should be interpreted cautiously. Another important limitation is the size of the synthetic datasets. Although DF2 and DF3 each contain 5,000 observations, these observations were algorithmically generated from only 137 original records. Therefore, they should not be treated as equivalent to 5,000 independently collected experimental observations. Together, these limitations reduce how much can be inferred from the regression results and reinforce the use of Level 3 mainly as a way to evaluate synthetic-data fidelity.

Given this information, future research should experiment using these synthetic methods on larger and more diverse LIB datasets to further observe whether trends remain consistent. In addition, studies should implement more complicated variables like internal resistance and battery chemistry in order to reinforce or contradict any particular findings. Another potential direction to take is to attempt to combine PADS with DDS into a hybrid model to observe the "grey area" between extreme plausibility and statistical fidelity. Finally, as stated earlier, multiple levels of evaluation should be considered when evaluating the legitimacy of synthetic data (19).

The objective of this study procedure was to compare two synthetic data generation methods (PADS and DDS) and ultimately determine their feasibility in real-world LIB research. Overall, the initial hypothesis was correct. The study's consensus is that DDS was the more effective method for this study; it had the greater ability to consistently preserve the original dataset's statistical trends and structure across three levels of evaluation. In contrast, the results for PADS had larger deviations, and thus were less effective, due to its emphasis on plausibility. Together, these findings suggest that DDS should be explored as a viable option for addressing battery-data scarcity, while PADS is better suited for other physically constrained applications.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Wen Cheng, Professor and Associate Chair at the Civil Engineering Department, California State Polytechnic University, Pomona, for his mentorship and guidance throughout this study.

FUNDING SOURCES

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. Kosonen I. Intermittency of renewable energy: review of current solutions and their sufficiency [Internet]. 2018 (accessed on 2026-06-21). Available from: http://lutpub.lut.fi/handle/10024/149296
2. International Energy Agency. Renewables 2023 [Internet]. 2023 (accessed on 2026-06-21). Available from: https://iea.blob.core.windows.net/assets/96d66a8b-d502-476b-ba94-54ffda84cf72/Renewables_2023.pdf
3. Kochtcheeva LV. Renewable energy: global challenges [Internet]. E-International Relations. 2016 May 27 (accessed on 2026-06-21). Available from: https://www.e-ir.info/2016/05/27/renewable-energy-global-challenges/
4. Surur FM, Mamo AA, Gebresilassie BG, Mekonen KA, Golda A, Behera RK, *et al*. Unlocking the power of machine learning in big data: a scoping survey. *Data Sci Manag.* 2025; 8 (4). doi:10.1016/j.dsm.2025.02.004.
5. Severson KA, Attia PM, Jin N, Perkins N, Jiang B, Yang Z, Chen MH, Aykol M, Herring PK, Fraggedakis D, Bazant MZ, Harris SJ, Chueh WC, Braatz RD.

Data-driven prediction of battery cycle life before capacity degradation. *Nat Energy*. 2019; 4 (5): 383-391. doi:10.1038/s41560-019-0356-8.

6. Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA); 2016. doi:10.1109/DSAA.2016.49.
7. Choi J, Choi J, Kang M. Generative modeling through the semi-dual formulation of unbalanced optimal transport. arXiv [Preprint]. 2023. doi:10.48550/arXiv.2305.14777.
8. Gonçalves PJ, Lueckmann JM, Deistler M, Nonnenmacher M, Öcal K, Bassetto G, *et al*. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*. 2020; 9: e56261. doi:10.7554/eLife.56261.
9. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J Comput Phys*. 2019; 378: 686-707. doi:10.1016/j.jcp.2018.10.045.
10. Richardson GW, Foster JM, Ranom R, Please CP, Ramos AM. Charge transport modelling of lithium-ion batteries. *Eur J Appl Math*. 2022; 33 (6): 983-1031. doi:10.1017/S0956792521000292.
11. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, *et al*. Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans Knowl Data Eng*. 2017; 29 (9): 2318-2331. doi:10.1109/TKDE.2017.2720168.
12. Santangelo G, Nicora G, Bellazzi R, Dagliati A. How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data. *BMC Med Inform Decis Mak*. 2025; 25: 89. doi:10.1186/s12911-024-02731-9.
13. Hernandez M, Osorio-Marulanda PA, Catalina M, Loinaz L, Epelde G, Aginako N. Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. *Front Digit Health*. 2025; 7: 1576290. doi:10.3389/fdgth.2025.1576290.
14. BatteryArchive.org [Internet]. 2026 (accessed on 2026-06-21). Available from: https://batteryarchive.org/cycle_list.html?time=0001
15. Manthiram A. An outlook on lithium ion battery technology. *ACS Cent Sci*. 2017; 3 (9): 1063-1069. doi:10.1021/acscentsci.7b00288.
16. Aslan E, Yasa Y. C-rate- and temperature-dependent state-of-charge estimation method for Li-ion batteries in electric vehicles. *Energies*. 2024; 17 (13): 3187. doi:10.3390/en17133187.
17. Wang H, Zhu Y, Kim SC, Pei A, Li Y, Boyle DT, *et al*. Underpotential lithium plating on graphite anodes caused by temperature heterogeneity. *Proc Natl Acad Sci U S A*. 2020; 117 (47): 29453-29461. doi:10.1073/pnas.2009221117.
18. An F, Zhang R, Wei Z, Li P. Multi-stage constant-current charging protocol for a high-energy-density pouch cell based on a 622NCM/graphite system. *RSC Adv*. 2019; 9 (37): 21498-21506. doi:10.1039/C9RA03629F.
19. Nowok B, Raab GM, Dibben C. synthpop: bespoke creation of synthetic data in R. *J Stat Softw*. 2016; 74 (10): 1-26. doi:10.18637/jss.v074.i11.
20. Li H, Xiong L, Jiang X. Differentially private synthesization of multi-dimensional data using copula functions. In: Proceedings of the 17th International Conference on Extending Database Technology; 2014 Mar 24-28; Athens, Greece. 2014. p. 475-486. doi:10.5441/002/edbt.2014.43.
21. NIST/SEMATECH. e-Handbook of Statistical Methods [Internet]. 2019 (accessed on 2026-06-21). Available from: https://doi.org/10.18434/M32189
22. Strang G. Lecture 8: norms of vectors and matrices [Internet]. In: Matrix methods in data analysis, signal processing, and machine learning. MIT OpenCourseWare; 2018 (accessed on 2026-06-21). Available from: https://ocw.mit.edu/courses/18-065-matrix-methods-in-data-analysis-signal-processing-and-machine-learning-spring-2018/resources/lecture-8-norms-of-vectors-and-matrices/
23. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. arXiv [Preprint]. 2019. doi:10.48550/arXiv.1907.00503.
24. Devie A, Baure G, Dubarry M. Intrinsic variability in the degradation of a batch of commercial 18650 lithium-ion cells. *Energies*. 2018; 11 (5): 1031. doi:10.3390/en11051031.
25. Juarez-Robles D, Jeevarajan JA, Mukherjee PP. Degradation-safety analytics in lithium-ion cells: part I. Aging under charge/discharge cycling. *J Electrochem Soc*. 2020; 167 (16): 160510. doi:10.1149/1945-7111/abc8c0.
26. Pathare A, Mangrulkar R, Suvarna K, Parekh A, Thakur G, Gawade A. Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *Int J Inf Manage Data Insights*. 2023; 3 (2): 100177. doi:10.1016/j.jjimei.2023.100177.

APPENDIX

Appendix A: Synthetic Generation Codes

A1.

Physics-Aware Data Synthesis (PADS) Code Snippets	
anode = [“graphite”] cathode = [“LFP”,”NMC”,”NCA”,”LCO”,”NMC-LCO”] source = [“snl”,”HNEI”,”calce”,”oxford”,”UL-PUR”] form = [“18650”,”pouch”,”prismatic”] Ah = [0.74, 1.1, 1.35, 2.8, 3.0, 3.2, 3.4]	Python lists were created to store allowable battery category ranges: (anode, cathode, source, form, Ah). During synthesis, each cell will have random properties sampled from these predefined values.

def soc_multiplier(max_soc, min_soc):
soc_range = max_soc - min_soc
# narrow range allows for more consistent results 	
factor2 = 1.2-0.003*(soc_range) #factor2 is kept within realistic limits
return np.clip(factor2, 0.8, 1.25)
	A custom-made state-of-charge (SOC) multiplier range was made to generate SOC rates that fall within the allowable range. Smaller ranges received higher scaling factors for balance. In addition, the “factor2” variable was bounded between 0.8 and 1.25 for realistic value generation.

if min_soc >= max_soc: min_soc = max(0, max_soc - rng.choice([20, 40]))	This code snippet prevents the minimum SOC from exceeding the maximum SOC by regenerating the minimum SOC to a constrained value.
The Python code displayed above are snippets from the Google Colab file titled “PADS Synthesis Generation Code”. For reproducibility purposes, the original dataset found in Appendix C will be needed. The full code is available via download below or by request, and must be opened on Google Colab: https://colab.research.google.com/drive/1q3h73-tX51C8XtLFWkVlf7PNGy06ZXbF?usp=drive_link	

A2.

Distributional Data Synthesis (DDS) Code Snippets	
continuous = ["Ah","Temperature (C)","Max SOC","Min SOC","Charge Rate (C)","Discharge Rate (C)"] all_cols = categories + continuous X = df1[all_cols].values.astype(float)	Base variables from the original dataset (Ah, Temperature, Max SOC, Min SOC, Charge Rate, Discharge Rate) are converted into a numerical matrix to later process.
U = ranks / (len(X) + 1) Z = np.sqrt(2) * special.erfinv(2*U-1)	This code references equation (6) and (7). First, the code changes the previously established rank observations into empirical uniform variables. Next, using equation (7), the code transforms these variables into latent Gaussian variables using inverse mapping.
M = 5000 Zz = np.random.multivariate_normal(mu, Sigma, M) Uu = 0.5 * (1 + special.erf(Zz / np.sqrt(2)))	This code references equation (8) and (8). By setting M equal to 5,000, the code utilizes equation (8) to generate the synthetic data using Gaussian copula functions. Afterwards, to restabilize the variables, the code uses equation (8) to convert the mapped variables back to the initial array.
The Python code displayed above are snippets from the Google Colab file titled "DDS Synthesis Generation Code". For reproducibility purposes, the original dataset found in Appendix C will be needed. The full code is available via download below or by request, and must be opened on Google Colab: https://colab.research.google.com/drive/1fV8bh-3g2eMdQ4QsHngFuEbHZXSMXZK-?usp=drive_link	

Appendix B: Evaluation Codes

B1.

Three-Level Synthetic Data Evaluation Framework (TSDEF) Code
This code's objective is to take .csv file input and produce three .csv files for output, with each spreadsheet listing quantitative results for each level. In order to run the program successfully, all three dataset .csv files must be uploaded. The full code is available via download below or by request, and must be opened on Google Colab. TSDEF Evaluation Systems Code

Appendix C: Raw Datasets

C1.

Raw Original Dataset (DF1)
The link below provides a Google Spreadsheet of all 137 original LIB data samples used for synthesis in this study. https://docs.google.com/spreadsheets/d/1l0jMY1Cje9P5DgaE6CHXYtoSLhpmXKlqphpZ8QfenhM/edit?gid=791660092#gid=791660092

C2.

Raw Physics-Aware Data Synthesis Dataset (PADS, DF2)
The link below provides a Google Spreadsheet of all 5,000 synthetic LIB data samples generated by PADS to form DF2 for this study. https://docs.google.com/spreadsheets/d/12VOOr0neovbJ7cmrVF0wCIuV5Lj-q2WXGYuGn0kBRcE/edit?gid=881483558#gid=881483558

C3.

Raw Distributional Data Synthesis Dataset (DDS, DF3)
The link below provides a Google Spreadsheet of all 5,000 synthetic LIB data samples generated by DDS to form DF3 for this study. https://docs.google.com/spreadsheets/d/1zBZ4egITgjWGYqHPlK7bpglfrUiE8_sFkrtVBZdhW1A/edit?gid=804935420#gid=804935420

Appendix D: Raw TSDEF Results

D1.

Raw Level 1 Results
The link below provides a Google Spreadsheet of the raw Level 1 results generated by the TSDEF Evaluator Code that was used in the "Results" section of the study. https://docs.google.com/spreadsheets/d/1XEZZtoHU3ko8E7EK9w4HmVpURrdBcQ_xR5RzvX6xYPU/edit?gid=682844635#gid=682844635

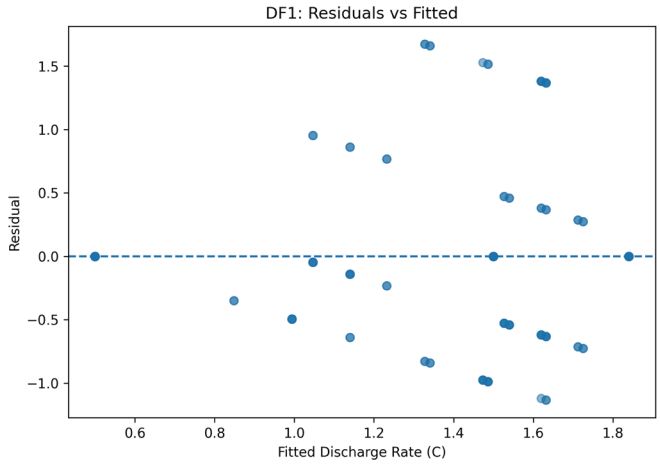
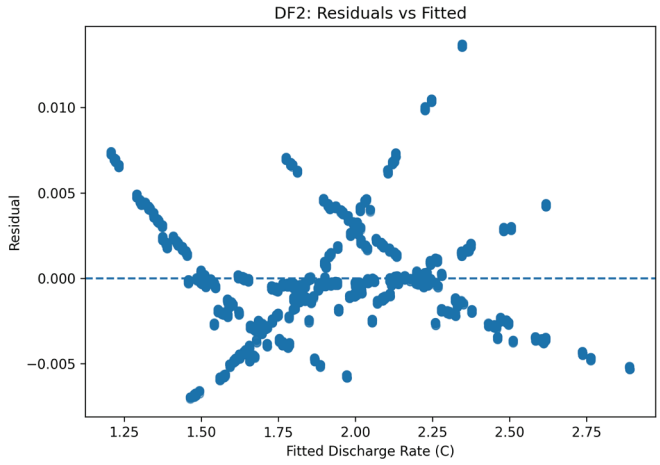
D2.

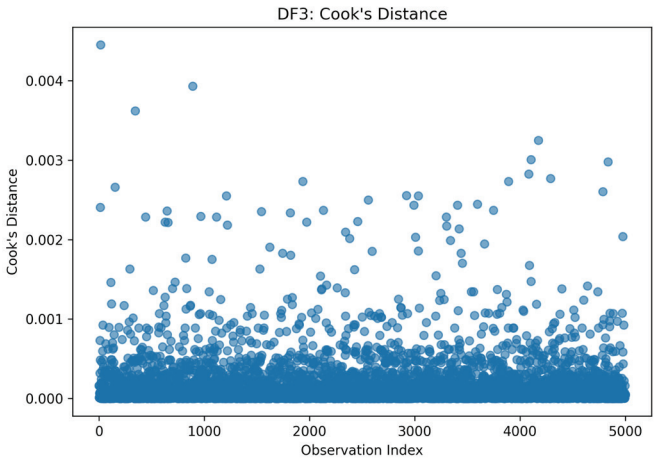
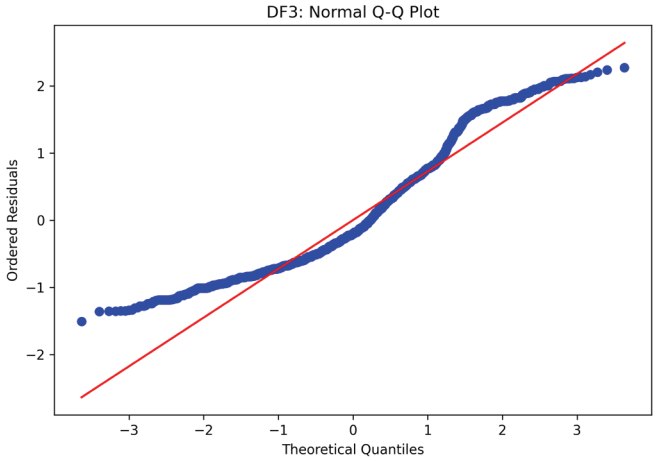
Raw Level 2 Results
The link below provides a Google Spreadsheet of the raw Level 2 results generated by the TSDEF Evaluator Code that was used in the "Results" section of the study. https://docs.google.com/spreadsheets/d/1Z0TsIC0XsiA4Q-S5nW5WHsqn5WjhX1PMkaW-AHYLtTQ/edit?gid=1170541472#gid=1170541472

D3.

Raw Level 3 Results
The link below provides a Google Spreadsheet of the raw Level 3 results generated by the TSDEF Evaluator Code that was used in the "Results" section of the study. The specific results are all in different tabs. https://docs.google.com/spreadsheets/d/1a363O6mV9LA9MXAgG3EldVIPj673u88fL_zru8rSS9U/edit?usp=sharing

D4.

Level 3 Data Graphs	
DF1: Residuals vs Fitted Graph	
DF2: Residuals vs Fitted Graph	

DF3: Cook's Distance Graph	
DF3: Q_Q Graph	
DF3: Residuals vs Fitted Graph