MDPI

Article

Stratified Fréchet Distance: A Three-Layer Diagnostic Framework for Conditional Time Series Generation under Data Scarcity

Tsuyoshi Okita[1]

[1] Kyushu Institute of Technology; tsuyoshi@ai.kyutech.ac.jp
* Correspondence: tsuyoshi@ai.kyutech.ac.jp

Abstract

Evaluating conditional time-series generation models remains challenging in battery research, where degradation data are often limited and experiments cover only a small number of operating conditions. The widely used Fréchet Inception Distance (FID) summarizes all conditions into a single score, which can obscure failures under rare but safety-critical conditions. Several condition-aware extensions of FID, including Conditional Fréchet Inception Distance (CFID), partially address this limitation by evaluating each condition separately. However, these approaches do not assess whether physically meaningful relationships between operating conditions are preserved, and their reliability deteriorates when only a few samples are available for each condition. To address these issues, we propose a three-layer diagnostic framework for evaluating conditional generative models under limited-data conditions. The first layer, Stratified Fréchet Distance, identifies the specific operating conditions and degradation phases where generation quality degrades. The second layer, based on Conditional Response Consistency (CRC), Conditional Distance Ratio (CDR), and Mean-Order Preservation (MOP), evaluates whether the model preserves the distance structure and ordering between conditions. MOP detects condition-ordering defects that CRC cannot identify when the real data distance matrix is non-monotone. This layer also enables statistically meaningful comparisons even when only a small number of samples are available. The third layer detects strata where statistical estimates are unreliable and provides a more stable alternative for evaluation. We validate the framework on four battery degradation datasets using two generative model architectures. The proposed approach reveals condition-specific failures that are not captured by conventional FID. It localizes generation errors to the late-stage high-temperature degradation regime that is most relevant to battery safety. The framework also detects structural distortions with statistical significance. In addition, it consistently ranks physics-informed model variants across quality differences spanning seven orders of magnitude. These results demonstrate that the proposed framework provides a practical and physically interpretable evaluation methodology for conditional generative modeling in battery degradation analysis.

Keywords: generative model evaluation, stratified evaluation, conditional response consistency, battery degradation, small-sample evaluation, time series generation

Received:
Revised:
Accepted:
Published:

1. Introduction

The safety of lithium-ion batteries depends critically on operating temperature. High temperatures accelerate SEI (solid electrolyte interphase, a passivation layer that forms on the anode surface) growth and cathode decomposition, increasing the risk of thermal runaway (uncontrolled exothermic reaction)[1]. Low temperatures promote lithium plating

(metallic lithium deposition on the anode) and dendrite-driven internal short circuits[2]. Quantitative characterization of degradation across all operating conditions is a prerequisite for safety assessment[3].

Yet experimental data under the extreme conditions most critical for safety—high or low temperature, high C-rate (the ratio of charge/discharge current to battery capacity)—are costly to acquire and inevitably underrepresented in existing datasets[4,5]. Deep generative models[6–9] have advanced to the point where synthesizing degradation curves for such conditions is becoming feasible[10–14]. In the emerging paradigm of battery digital twins[15,16], synthetic data are already used to augment limited experiments and simulate untested scenarios. A digital twin maintains a virtual replica of a physical battery system, continuously updated with sensor data, and uses generative models to predict degradation under conditions not yet encountered in service—for example, forecasting capacity fade at temperatures outside the tested range or under novel charge protocols. The fidelity of such predictions depends directly on the quality of the underlying generative model at each operating condition. As safety decisions increasingly rely on such data, their trustworthiness becomes a practical engineering concern.

Evaluating synthetic degradation data, however, requires answering two distinct questions that current metrics conflate or ignore. The first is *per-condition quality*: does the generative model produce accurate curves at each temperature and in each degradation phase? The second is *inter-condition structural consistency*: does the model preserve the physical relationships across conditions—for example, the monotonic acceleration of degradation from 25°C to 43°C?

These questions are difficult to answer because battery degradation data possess three characteristics that set them apart from the large-scale image datasets for which evaluation metrics such as FID were designed.

First, the data are scarce. Battery datasets typically contain tens to at most a few hundred cells[4,5], in contrast to the tens of thousands of samples available for image generation. Per-condition evaluation must therefore work reliably with extremely limited samples—a regime in which covariance-based metrics become statistically unstable.

Second, physical conditions induce qualitative changes in the degradation trajectory. Changing the temperature from 25°C to 43°C changes the dominant degradation mechanism (e.g., from gradual SEI thickening to rapid electrolyte decomposition), producing curves with qualitatively different shapes and lifetimes[3]. Similar condition-dependent changes arise in other physics-governed domains, such as rotating machinery fault diagnosis[17] and materials processing[18].

Third, the most important conditions are the rarest. Extreme temperatures and high C-rates are most relevant to safety yet most expensive to test. Mayer et al.[19] have shown that generative models trained by maximum likelihood inherently under-represent such minorities, making quality evaluation under these conditions doubly important.

The standard metric, Fréchet Inception Distance (FID)[20], addresses none of these challenges. It aggregates all data into a single scalar, and its feature space has been shown to mask quality differences even when they are perceptually apparent[21]. CFID[22] and Fréchet Joint Distance (FJD)[23] evaluate per-condition quality for discrete image classes. Neither addresses temporal diagnosis or confounding detection, which require stratification along additional axes. Neither assesses whether the generative model preserves the physical relationships between conditions—a question that calls for a dedicated structural verification step. And neither accounts for the statistical unreliability of covariance-based estimates when only a handful of samples are available per condition.

Consider a model that generates good 25°C curves but poor 43°C curves (12% of the dataset; Fig. 1). FID dilutes the 43°C failure into the 88% of correct samples and remains

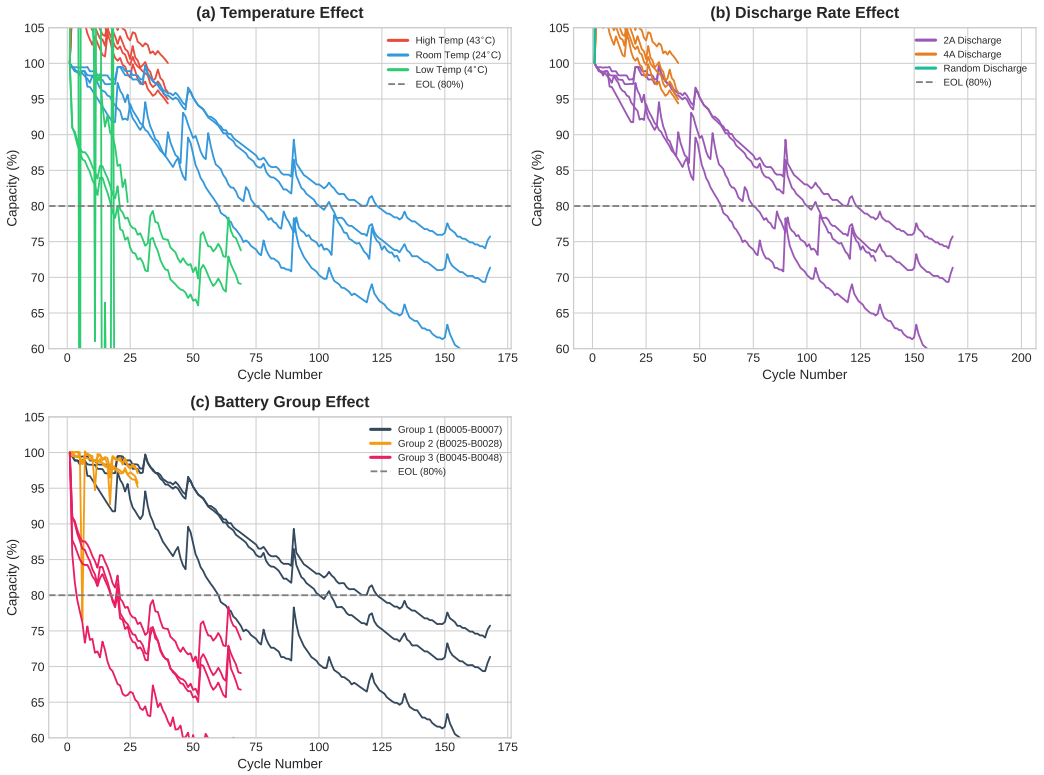

Figure 1. Temperature-dependent capacity degradation curves from the NASA battery dataset[24]. Degradation rate and pattern vary substantially across temperatures. Accurate reproduction of minority-condition behavior (43°C, 44°C) is a prerequisite for safety evaluation. The remaining three datasets (SNL, MICH, CALB) used in this study are from the BatteryLife benchmark[25].

virtually unchanged. Dilution also occurs along the temporal axis: battery degradation proceeds through linear fade followed by nonlinear acceleration beyond the knee point (the inflection where capacity fade transitions from gradual to rapid)[26], and FID lets early-phase success mask late-phase failure. Even if a model generates plausible curves at each temperature individually, it may distort the relative distances between conditions—producing 43°C curves that look more like 25°C than they should. Neither FID nor CFID can detect this structural distortion.

To address these challenges, we propose a three-layer diagnostic framework. The first layer, **Stratified Fréchet Distance (SFD)**, partitions data by operating condition, temporal segment, or both, and diagnoses *where* and *when* generation quality breaks down. SFD with $\lambda = 0$ reduces exactly to CFID and is adopted as the default. Inter-condition consistency, previously addressed by the Between-SFD term ($\lambda > 0$), is now handled more directly by the second layer. The second layer introduces two new metrics for inter-condition structural verification: **Conditional Response Consistency (CRC)** for datasets with three or more conditions, and **Conditional Distance Ratio (CDR)** for datasets with only two. Both measure whether the generative model preserves the distance structure between conditions, and permutation tests provide statistical significance. The third layer addresses reliability under data scarcity: **Effective Stratum Size (ESS)** flags strata where Fréchet distance estimation is unreliable, and **Stratified MMD (S-MMD)** provides a covariance-free alternative for such strata.

We validate the framework on four battery datasets (NASA[24], SNL, MICH, CALB[25], 161 cells in total) with conditional VAE and conditional Flow Matching[27] models. The experiments reveal that SFD identifies per-condition quality degradation that FID entirely misses. In a controlled experiment where a model is trained on a single tem-

perature, FID remains virtually unchanged from the baseline, whereas SFD reveals that the excluded conditions are nearly twice as far from reality as the training condition. Condition-by-time stratification further localizes the largest quality gap to the late degradation phase at 35°C, the regime most relevant to safety. The structural verification layer confirms that well-trained generators preserve the physical ordering of inter-condition distances while deficient generators distort it. A CRC permutation test on cross-stratified condition pairs achieves statistical significance ($p = 0.001$) for temperature–C-rate structured data. MOP detects condition-ordering defects across four battery datasets with 100% detection rate on three datasets ($p < 0.0001$, $n_{\text{seeds}} = 10$). CDR extends this structural verification to datasets with only two conditions, reliably distinguishing normal from distorted generation with bootstrap confidence intervals. These findings hold across all four datasets and across two fundamentally different generative architectures (CVAE and conditional Flow Matching with Fourier Neural Operator).

This work makes four contributions. First, we characterize three properties of physics-governed time series generation—data scarcity, condition-dependent qualitative changes, and minority-condition importance—that make evaluation fundamentally different from large-scale image generation and that existing metrics do not address. Second, we propose a three-layer diagnostic framework comprising SFD for per-condition quality diagnosis; CRC, CDR, and MOP for inter-condition structural verification with statistical significance; and ESS and S-MMD for small-sample reliability assessment. CRC and CDR are motivated by Benny et al.'s between-class evaluation [28] and extend it by testing rank ordering and distance magnitude preservation with formal permutation tests. MOP is a new indicator that detects condition-ordering defects (type-2 inversion and type-3 confusion) when CRC lacks power due to small K or non-monotone real data. Third, we analyze the detection power of Fréchet distance and MMD in the small-sample regime, derive the optimal feature dimension $d^* = O(n_s^{1/4})$, and establish the complementarity regime in which MMD-based evaluation becomes preferable to FD-based evaluation. Fourth, we validate the framework through fourteen experiments across four datasets, three feature extractors, and two generative architectures, confirming that the diagnostic capabilities are robust to the choice of data source, feature space, and generative model. Experiments 1–9 validate Layer 1. Experiment 10 introduces and validates MOP for Layer 2. Experiments 11–14, reported in Appendix C, provide detailed Layer 2 validation covering seed dependence, cross-dataset generalization, CRC minimum-K requirements, and CDR inflation conditions. The experiments demonstrate increased diagnostic sensitivity under controlled failure scenarios, rather than an improvement in generative model performance.

The remainder of this paper is structured as follows. Section II reviews the background on conditional time series generation and formalizes the dilution problem. Section III presents the three-layer diagnostic framework in detail: Layer 1 (SFD and its stratification variants), Layer 2 (CRC and CDR for structural verification), and Layer 3 (ESS and S-MMD for reliability assessment). Sections IV and V describe the experimental setup and report results. Section VI discusses the implications of the findings and their limitations. Section VII concludes the paper.

2. Background

2.1. Conditional Time Series Generation

A conditional time series generation model receives a condition parameter c (e.g., temperature or C-rate) and learns the conditional distribution $p(x|c)$ of the time series x. The paradigm originated in image synthesis with Conditional GANs[29] and Conditional VAEs[30], and has since been extended through classifier-free guidance[31] and Conditional Flow Matching[9].

Adaptation to time series has proceeded along several lines. TimeGAN[11] introduced an embedding network to capture temporal dynamics. TimeVAE[12] proposed a VAE architecture tailored to sequential structure. CSDI[13] applied conditional score-based diffusion to time series imputation. In the battery domain, DiffBatt[14] demonstrated that conditional generation of degradation curves can be combined with lifetime prediction for practical data augmentation.

These advances make it increasingly feasible to synthesize degradation curves under conditions for which no experimental data exist. Beyond data augmentation, conditional generation enables scenario analysis—synthesizing degradation trajectories under hypothetical operating conditions (e.g., a temperature profile not tested in the laboratory) to support design decisions and safety certification[1]. The value of such scenario generation depends entirely on the fidelity of the generated data at each target condition, making per-condition quality evaluation a prerequisite for any downstream use.

2.2. The Fréchet Inception Distance and Its Limitations

FID[20] is the standard metric for evaluating generative models. It fits Gaussian distributions to the features of real and generated data and computes their Fréchet distance:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{1}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of the real and generated feature distributions, respectively.

FID models all data as a single Gaussian. It does not distinguish which operating condition a sample belongs to or which temporal segment of a time series it represents. For homogeneous data this approximation is acceptable. For battery degradation curves, where distributions differ qualitatively across temperatures, fitting a single Gaussian to the mixture is itself a poor modeling choice.

When FID is applied to time series, domain-specific feature extractors such as InceptionTime[32] replace the Inception network used in image generation[33]. Changing the feature extractor, however, does not alter the structural limitation: FID aggregates a mixture distribution into a single scalar. Stein et al.[21] have further shown that the Inception-V3 feature space can fail to reflect the perceptual quality of generated samples. The limitation we address is more fundamental: it originates not in the choice of feature extractor but in the manner of aggregation itself.

Several conditional evaluation metrics have been proposed. CFID[22] computes FID separately for each class label and averages the results. FJD[23] evaluates the joint distribution of images and conditions. FPD[34] targets physical simulation data. All of these assume discrete class labels (e.g., “dog” or “cat”) and large sample sizes per class. Their extension to continuous physical parameters, temporal quality diagnosis, inter-condition structural verification, confounding detection, and the small-sample regime has not been investigated.

2.3. The Dilution Problem

To formalize FID’s limitation, consider K conditions $\{c_1, \ldots, c_K\}$ with data proportions $p(c_k)$. Suppose a generative model fails only under condition c^* while producing perfect output elsewhere. Because FID evaluates the mixture distribution, the contribution of the failed condition is weighted by its proportion:

$$\text{FID} \approx p(c^*) \cdot d_F(p(x|c^*), q(x|c^*)) \tag{2}$$

Two consequences follow. First, when $p(c^*) = 0.12$ (e.g., 43°C constitutes 12% of the dataset), a complete failure at that condition contributes only 12% to FID: this is *condition-axis dilution*. The rarer a condition, the more severely its failures are diluted—yet rare conditions are precisely those most relevant to safety.

Second, dilution is not confined to the condition axis. Along the temporal axis, a failure in the late degradation phase can be diluted by success in the early phase. When multiple condition parameters are confounded (e.g., temperature and C-rate), evaluating by one variable alone allows the effect of the other to leak into the assessment.

All of these problems stem from pooling subsets with fundamentally different characteristics and evaluating them without distinction. In classical statistics, stratified sampling addresses exactly this situation by respecting the internal structure of the population. The three-layer diagnostic framework proposed in the next section applies this principle to generative model evaluation. Layer 1 (SFD) resolves dilution through stratification. Layer 2 (CRC and CDR) addresses inter-condition structural verification. Layer 3 (ESS and S-MMD) manages statistical reliability under data scarcity.

3. Proposed Diagnostic Framework for Conditional Generation Quality

Evaluating conditional generation quality under data scarcity requires answering three distinct questions, each demanding its own analytical tool. The first question is *where and when* quality breaks down across operating conditions and degradation phases. The second is *whether* the generative model preserves the physical relationships that link one condition to another—for instance, the monotonic acceleration of degradation from moderate to high temperatures. The third is *how reliable* the per-stratum estimates are when some conditions contain as few as three or four batteries.

We organize the framework into three corresponding layers (Table 1). Layer 1 provides per-condition and per-time-segment quality diagnosis through Stratified Fréchet Distance (SFD). Layer 2 introduces Conditional Response Consistency (CRC) and Conditional Distance Ratio (CDR) to verify that the inter-condition distance structure is preserved, with permutation tests [35] supplying statistical significance. Mean-Order Preservation (MOP) is also introduced in Layer 2 to detect condition-ordering defects (type-2 inversion and type-3 confusion) when CRC lacks statistical power due to small K or non-monotone real data distance matrices. Validation is reported in Experiment 10 (Section V) and Experiments 11–14 (Appendix C). Layer 3 equips the practitioner with Effective Stratum Size (ESS) and Stratified MMD (S-MMD) to assess estimation reliability and to obtain stable alternatives when Fréchet distance estimates are unreliable.

The three layers are designed to be used in sequence. A practitioner begins by computing SFD to identify problematic conditions and time segments (Layer 1). The next step is to apply CRC or CDR to determine whether the inter-condition structure is intact (Layer 2). Finally, ESS and S-MMD are consulted to judge which Layer 1 estimates can be trusted and which require caution (Layer 3). If the permutation test in Layer 2 fails to achieve significance because the number of conditions K is small, cross-stratification along a second axis (e.g., temperature$\times$time) increases the number of condition pairs and thereby improves statistical power. If per-stratum sample sizes are below ten, mean-vector Euclidean distance should replace MMD in the Layer 2 computation to avoid kernel bandwidth instability.

3.1. Layer 1: Quality Diagnosis — Stratified Fréchet Distance (SFD)

The root cause of the dilution problem identified in Section II is that FID computes a single distance over the pooled data, allowing failures at minority conditions to be absorbed

Table 1. Overview of the diagnostic framework.

Metric	What it measures	Applicable when	Key output	Experiments
Layer 1: Quality Diagnosis				
SFD_c	Per-condition quality	$K \geq 2$ conditions	FD per condition	1–6
SFD_t	Per-time-segment quality	Time-series data	FD per segment	7, 7b
$SFD_{c\times t}$	Condition×time quality	$K \geq 2$, time-series	2D quality map	7, 7b
CI / SFD_{joint}	Confounding detection	Two granularities	Ratio > 1 = confound	8, 8b
Layer 2: Structure Verification				
CRC	Distance structure preservation	$K \geq 3$ ($\geq$3 pairs)	ρ_S + p-value	1, 2, 7, 8, 9, 13
CDR	Distance structure preservation	$K = 2$ (1 pair)	Ratio + bootstrap CI	1, 2, 7, 9, 14
MOP	Condition ordering preservation	$K \geq 2$, $\rho_{proj} \geq 0.7$	Spearman ρ (< 0 = inversion)	10–14
Layer 3: Reliability Assessment				
ESS	Estimation reliability	Any stratum	< 1 = unreliable	1, 2, 7, 9
S-MMD	Covariance-free distance	ESS < 1 strata	MMD per stratum	1, 2, 7, 9
ConsI	SFD–S-MMD agreement	$K \geq 3$ strata	Spearman ρ	1, 2

by the majority. SFD resolves this by partitioning data into strata according to a stratification variable s and computing the Fréchet distance independently within each stratum:

$$\mathrm{SFD}(s) = \underbrace{\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} d_F(p(x|s), q(x|s))}_{\text{Within-SFD}} + \lambda \cdot \underbrace{d_F(p(x), q(x))}_{\text{Between-SFD}} \tag{3}$$

where $\mathcal{S}$ is the set of strata, d_F denotes the Fréchet distance, and $\lambda \geq 0$ controls the weight of the overall distributional consistency term. Within-SFD assigns equal weight to every stratum, ensuring that a minority condition comprising 6% of the data receives the same attention as one comprising 40%. This equal-weight design is deliberate. A failure at 43°C is no less consequential than a failure at 4°C simply because fewer batteries were tested at that temperature. Weighting each stratum by its sample proportion $p(c_k)$ would reproduce precisely the dilution that SFD is designed to avoid. When domain knowledge suggests that certain conditions are more important than others, a weighted average with user-specified weights w_k can be substituted without altering the framework. When $\lambda = 0$, the formulation reduces exactly to CFID[22]. Between-SFD (the $\lambda > 0$ term) is the Fréchet distance over the entire mixture distribution, identical to FID. Setting $\lambda > 0$ previously served to assess inter-condition distributional consistency alongside per-condition quality. In the present framework, this role is handled more directly and with statistical significance by Layer 2 (CRC and CDR), and $\lambda = 0$ is therefore adopted as the recommended default.

The choice of stratification variable determines the diagnostic question that SFD answers (Table 2). Stratification by operating condition (SFD_c) reveals which conditions exhibit poor generation quality. Stratification by temporal segment (SFD_t) reveals in which phase of the degradation curve the model fails, distinguishing, for example, early linear fade from late nonlinear acceleration beyond the knee point[26]. A failure pattern in which

Table 2. Stratification variables of SFD and their diagnostic roles.

Stratification s	Name	Diagnostic question	Prior work	Experiments
None (pooled)	FID	Overall quality?	FID[20]	1–8
Condition c	SFD_c	Which conditions fail?	CFID[22]	1–6
Time segment k	SFD_t	Which time window fails?	(novel)	7, 7b
(c, k)	$\text{SFD}_{c\times t}$	Condition × time?	(novel)	7, 7b
(c_1, c_2)	$\text{SFD}_{\text{joint}}$	Confounding?	(novel)	8, 8b

the generative model reproduces the early phase correctly but fails in the late phase is difficult to detect with FID or SFD_c, both of which compute features over the entire curve. SFD_t detects this failure directly as an elevated Fréchet distance in the late-segment stratum. Stratification by condition×time ($\text{SFD}_{c\times t}$) provides the finest granularity and can localize a failure to a specific condition in a specific degradation phase. Stratification by multiple condition parameters ($\text{SFD}_{\text{joint}}$) exposes inter-condition confounding.

To quantify confounding between two conditioning variables, we define the Confounding Index (CI) as the ratio of SFD computed at a finer stratification granularity to SFD at a coarser one:

$$\text{CI} = \frac{\text{SFD}_{\text{joint}}(c_1, c_2)}{\text{SFD}_{\text{single}}(c_1)} \tag{4}$$

A value of CI close to unity indicates that adding the second variable does not materially change the evaluation, implying little confounding. A value exceeding unity indicates that the finer stratification has exposed quality problems hidden at the coarser level, providing a concrete recommendation for which variables to include in model conditioning. CI thus serves not only as a post-hoc evaluation tool but also as an upstream design guide. A CI substantially above unity signals that the omitted conditioning variable should be incorporated into the generative model.

3.2. Layer 2: Structure Verification — CRC and CDR

Layer 1 diagnoses quality within each condition but does not assess whether the generative model preserves the relative distances between conditions. A model might produce plausible curves at each temperature yet compress or invert the distance between 25°C and 43°C, a distortion invisible to any per-condition metric. Layer 2 addresses this gap.

For each pair of conditions (c_i, c_j), define the inter-condition distance in the real data as $\Delta_{\text{real}}(c_i, c_j) = d(p(x|c_i), p(x|c_j))$ and in the generated data as $\Delta_{\text{gen}}(c_i, c_j) = d(q(x|c_i), q(x|c_j))$. When three or more conditions are available ($K \geq 3$), we measure the structural consistency of these distances via the Conditional Response Consistency:

$$\text{CRC} = \rho_S\big(\text{vec}(\mathbf{D}_{\text{real}}),\ \text{vec}(\mathbf{D}_{\text{gen}})\big) \tag{5}$$

where ρ_S is Spearman's rank correlation [36] and $\text{vec}(\cdot)$ extracts the $\binom{K}{2}$ upper-triangular entries of the pairwise distance matrix. CRC equal to unity means the ordering of inter-condition distances is perfectly preserved. CRC near zero means the structure has been scrambled. Negative CRC indicates an inversion.

When only two conditions are available, CRC cannot be computed because a single pair provides no basis for correlation. For this case we define the Conditional Distance Ratio:

$$\text{CDR}(c_1, c_2) = \frac{d(q(x|c_1),\ q(x|c_2))}{d(p(x|c_1),\ p(x|c_2))} \tag{6}$$

CDR near unity indicates that the inter-condition distance is preserved. A bootstrap confidence interval [37] that excludes unity signals statistically significant distortion. Temporal cross-stratification can increase the effective number of conditions (e.g., splitting two temperatures into first-half and second-half yields four cross-conditions and six pairs), enabling CRC computation even for two-condition datasets.

When permutation tests on CRC or CDR fail to reach significance because K is small, cross-stratification along a second axis increases the number of condition pairs and improves statistical power. In our experiments, cross-stratification by temperature$\times$time improves p from 0.067 ($K = 3$, 3 pairs) to 0.001 ($K = 6$, 15 pairs).

The inter-condition distance d can be any distributional metric. When per-condition samples are few ($n < 10$), we recommend the Euclidean distance between mean feature vectors, which requires no covariance or kernel estimation and remains stable even at $n = 3$. When samples are more plentiful ($n \geq 10$), MMD captures richer distributional differences. In our experiments, switching from MMD to mean-vector distance improved CRC on the NASA dataset from 0.44 to 0.95.

Mean-Order Preservation (MOP). CRC measures consistency of inter-condition distances but cannot detect type-2 inversion defects when the real data distance matrix is itself non-monotone—a common occurrence in battery datasets where individual cell variability can dominate temperature effects. MOP addresses this limitation by measuring whether the generative model preserves the *ordering* of conditions along the conditioning axis.

Let $\mathbf{v}_1$ be the direction in feature space maximally correlated with the temperature label (obtained by regressing the condition-wise mean vectors onto the normalized temperature). Define projected centroids $\pi_k^{\text{real}} = (\bar{\mathbf{x}}_k^{\text{real}} - \bar{\mathbf{x}}) \cdot \mathbf{v}_1$ and $\pi_k^{\text{gen}} = (\bar{\mathbf{x}}_k^{\text{gen}} - \bar{\mathbf{x}}) \cdot \mathbf{v}_1$, where $\bar{\mathbf{x}}$ is the global mean of the real data. The MOP score is defined as

$$\text{MOP} = \rho_S\Big(\{\pi_k^{\text{real}}\}_{k=1}^K,\ \{\pi_k^{\text{gen}}\}_{k=1}^K\Big). \tag{7}$$

MOP near $+1$ indicates that the generator preserves the physical ordering of conditions. MOP near -1 indicates that the ordering is inverted.

Statistical significance is assessed by a Mann-Whitney U test [38] comparing MOP distributions of the candidate model ($n_{\text{seeds}} \geq 5$ independent runs) against a correctly conditioned reference. The detection thresholds are

- **Type-2 inversion**: MOP < 0 ($p < 0.05$, Mann-Whitney).
- **Type-3 confusion**: MOP $< \mu_{\text{normal}} - 0.5$, where μ_{normal} is the reference MOP mean.

The projection quality $\rho_{\text{proj}} = \rho_S(\{\pi_k^{\text{real}}\}, \{t_k\})$ (Spearman correlation with the normalized temperature label t_k) must satisfy $\rho_{\text{proj}} \geq 0.7$ for reliable MOP detection.

3.3. Layer 3: Reliability Assessment — ESS and S-MMD

Computing the Fréchet distance requires estimating a covariance matrix for each stratum, and this estimate becomes unreliable when the number of samples is smaller than the feature dimension. Layer 3 provides two tools for managing this limitation.

Effective Stratum Size, defined as $\text{ESS}(s_k) = n_k/(d+1)$, expresses the ratio of per-stratum sample size n_k to the number of parameters needed for a full-rank covariance estimate. ESS below one indicates rank deficiency and unreliable Fréchet distance. ESS between one and three is marginally adequate. ESS above three suggests reliable estimation.

For strata with low ESS, Stratified MMD provides a covariance-free alternative. MMD[39] measures distributional distance through pairwise kernel evaluations and remains stable even at $n = 2$–3. S-MMD applies the same stratification structure as SFD:

$$\text{S-MMD}(s) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \text{MMD}^2(p(x|s), q(x|s)) \tag{8}$$

The Consistency Index (ConsI), defined as the Spearman rank correlation between per-stratum SFD and S-MMD values, quantifies whether the two metrics agree on which strata are problematic. High ConsI corroborates the SFD diagnosis. Low ConsI warns that estimation noise may be distorting the Fréchet distance in some strata.

A theoretical analysis (Appendix A) formalizes the complementarity between the two metrics. The minimum detectable distributional difference scales as $\Omega(d/\sqrt{n_s})$ for the Fréchet distance but only $\Omega(1/\sqrt{n_s})$ for MMD. The resulting $\Omega(d)$-fold gap implies that, for the 17-dimensional features used in our experiments, the Fréchet distance requires differences at least an order of magnitude larger than MMD to achieve comparable detection power. The analysis also yields the guideline $d^* = O(n_s^{1/4})$ for adapting the feature dimension to the per-stratum sample size.

3.4. Implementation

All metrics in the framework operate on feature vectors extracted from the raw degradation curves. To verify that results are not artifacts of a particular feature space, we compare three extractors: hand-crafted degradation features (17 dimensions, comprising segment-wise rates and first-difference statistics), an InceptionTime-style 1D-CNN (32 dimensions, trained on a temperature classification proxy task), and a temporal autoencoder (16 dimensions, trained on unsupervised reconstruction). The hand-crafted features capture degradation rate and curve shape from domain knowledge. The InceptionTime-style CNN uses multi-scale convolutions with residual connections trained on temperature classification, yielding a feature space that emphasizes inter-condition differences. The temporal autoencoder is trained on unsupervised reconstruction and captures the intrinsic structure of the data without requiring condition labels.

The SFD algorithm partitions data by stratum, computes the Fréchet distance within each, and returns their weighted sum (pseudocode in Appendix B). Its computational cost is $O(nd^2 + |\mathcal{S}| \cdot d^3)$, which is negligible relative to FID when the number of strata remains modest. CRC and CDR require $\binom{K}{2}$ pairwise distance computations, each costing $O(n_k d)$ for the mean-vector metric or $O(n_k^2)$ for MMD. S-MMD (Layer 3) applies the same stratification structure as SFD and incurs a cost of $O(n_k^2)$ per stratum for kernel evaluations. For the battery datasets used in this study ($|\mathcal{S}| \leq 6$, $n_k \leq 34$), this cost is negligible.

4. Experimental Setup

We validate the framework on four battery degradation datasets totaling 161 cells. This section describes the datasets, preprocessing, and generative model configuration. The experimental design and results are presented in Section V. The experiments are organized to reflect the layered structure of the framework. They progress from condition-axis stratification (Experiments 1–2) through feature extractor robustness (Experiment 6), temporal-axis stratification (Experiment 7), and confounding detection (Experiment 8), to validation on an alternative generative architecture (Experiment 9). This ordering follows the diagnostic workflow of Section III: Layer 1 is established first, then Layer 2 and Layer 3 assessments are introduced alongside each experiment.

Table 3. NASA battery dataset composition.

Temperature	Batteries	Proportion
4°C	13	39.4%
22°C	2	6.1%
24°C	11	33.3%
43°C	4	12.1%
44°C	3	9.1%
Total	33	100%

Table 4. BatteryLife dataset composition (number of cells by temperature).

	0°C	15°C	25°C	35°C	45°C	Total
SNL	–	9	34	18	–	61
MICH	–	–	21	–	19	40
CALB	8	–	2	14	3	27

4.1. Datasets

The **NASA Battery Dataset**[24] comprises capacity degradation curves from 33 lithium-ion cells tested at five temperatures (Table 3). The pronounced imbalance—two cells at 22°C versus thirteen at 4°C—makes this dataset well suited for testing SFD's ability to detect minority-condition failures that FID would dilute.

The **BatteryLife Dataset**[25] integrates voltage-profile data from 128 cells collected at three institutions: SNL (Sandia National Laboratories, 61 cells), MICH (University of Michigan, 40 cells), and CALB (China Aviation Lithium Battery, 27 cells), as summarized in Table 4. Mean normalized voltage per cycle serves as a proxy for capacity degradation. The SNL subset is of particular interest because it spans three temperatures and four C-rates. The 0.5C and 3C rates, however, were tested only at 25°C (Table 5), introducing a temperature–C-rate confound that we exploit in Experiment 8.

4.2. Preprocessing

Each battery's full degradation curve constitutes a single sample. This granularity corresponds to the application scenario of synthesizing degradation trajectories under unobserved conditions from a small number of experimentally tested cells, and differs from cycle-level prediction tasks in which individual charge–discharge cycles serve as data points. All curves were interpolated to 50 equally spaced points and normalized by their initial values. A 17-dimensional hand-crafted feature vector was then extracted from each curve and standardized.

4.3. CVAE Models

Experiment 1 uses simulated generators (condition confusion) to establish the detection principle in a controlled setting. From Experiment 2 onward, we validate the framework using a Conditional VAE rather than simulated generators. Experiment 9 further extends the validation to a third generative architecture (HPC-FNO-CFM[27], conditional Flow Matching with Fourier Neural Operator), confirming that the diagnostic capabilities generalize beyond CVAE. The CVAE comprises a two-layer encoder and decoder (hidden dimension 64, latent dimension 8) trained for 1500 epochs with Adam ($lr = 10^{-3}$) on the SNL subset (61 cells, 3 temperatures). To create controlled quality degradations, we train four model variants by progressively excluding temperature conditions from the training set. Model A (baseline) uses all three temperatures. Model B excludes the minority condition 15°C (14.8% of data). Model C retains the same architecture as Model A but

Table 5. Temperature×C-rate structure of the SNL subset. Empty cells indicate untested combinations, constituting a confound between temperature and C-rate.

	0.5C	1C	2C	3C
15°C	–	4	5	–
25°C	8	12	6	8
35°C	–	12	6	–

swaps the 15°C label with 25°C at generation time. Model D excludes both 15°C and 35°C, training on 25°C data only. Model D represents the most extreme case of condition omission, deliberately engineered to test whether FID is capable of detecting the resulting quality gap. The central question is: when a model has been trained on a single temperature but is asked to generate curves for all temperatures, does FID flag the problem—or does it remain silent? Unless otherwise stated, results for CVAE-based experiments (Experiments 2–7) are reported as means over three random seeds, with bootstrap 95% confidence intervals (500 iterations) to quantify estimation uncertainty. This choice accounts for the small per-stratum sample sizes that make seed-to-seed variance an insufficient indicator of reliability. Experiment 8 uses $n_{\text{seeds}} = 10$ with a Mann-Whitney U test to assess whether Model T consistently exceeds Model J across conditions; the increased seed count reflects the distributional nature of this comparison rather than a point-estimation task. Experiments 10–12 also use $n_{\text{seeds}} = 10$; Experiment 14 uses $n_{\text{seeds}} = 5$. Experiment 1 uses fixed simulated generators and Experiment 9 evaluates a pre-trained FNO architecture; random seed variation does not apply to these experiments.

4.4. Computational Environment and Reproducibility

All experiments were conducted on a Linux workstation equipped with an NVIDIA RTX 3090 GPU (24 GB VRAM) and an Intel Xeon processor. The framework is implemented in Python 3.10 using PyTorch 2.0, NumPy, SciPy, and scikit-learn. CVAE training for a single model variant completes in approximately 5 minutes. SFD/CRC/CDR evaluation of all experiments requires approximately 30 minutes including bootstrap confidence intervals (500 iterations). Source code for the framework and all experiments will be made publicly available upon acceptance.

5. Results

We report results from fourteen experiments organized by the three layers of the diagnostic framework. *Layer 1 experiments (Experiments 1–9)* validate SFD and its variants. Seven are presented in Section 5. Experiments 1 and 2 establish the basic detection capability of SFD using condition-axis stratification; Experiment 6 verifies robustness across three feature extractors; Experiment 7 introduces temporal and condition×time stratification; Experiment 8 demonstrates confounding detection via temperature–C-rate stratification; Experiment 9 validates the framework on an alternative generative architecture (HPC-FNO-CFM, conditional Flow Matching with Fourier Neural Operator). Five Layer 1 experiments are reported in Appendix C: Experiments 3–5 examine λ sensitivity, cross-dataset generalization, and the formal relationship with CFID; Experiments 7b and 8b extend Experiments 7 and 8 with alternative temporal segmentation and a second confounding source, respectively.

Layer 2 experiments (Experiments 10–14) validate MOP, CRC, and CDR. Experiment 10 (Section 5.2) establishes MOP as a reliable detector of condition-ordering defects across generator types. Experiments 11–14, reported in Appendix C, provide further validation. Experiment 11 examines seed dependence, Experiment 12 examines cross-dataset general-

Table 6. Condition confusion simulation (NASA). DA quantifies the sensitivity gain of SFD over FID for each confusion pattern.

Confusion pattern	Proportion	FID ratio	FD ratio	DA
22°C→24°C	6.1%	2.74	215.9	78.8
22°C→44°C	6.1%	82.2	5287.6	64.3
43°C→24°C	12.1%	6.41	192.4	30.0
43°C→4°C	12.1%	95.2	2344.4	24.6
4°C→24°C	39.4%	2700.1	6629.9	2.46

ization of MOP, Experiment 13 examines the minimum K required for CRC significance, and Experiment 14 identifies the root cause of CDR inflation under data scarcity. CRC, CDR, MOP, and S-MMD are computed in parallel with all experiments to provide Layer 2 and Layer 3 assessments alongside the Layer 1 results.

The three layers illustrate a diagnostic workflow. A practitioner first identifies *where* quality breaks down by condition and degradation phase (Layer 1), then tests *whether the physical relationships across conditions are intact* (Layer 2), and finally assesses *how much to trust each per-stratum estimate* (Layer 3).

5.1. Layer 1: Per-Condition Quality Diagnosis (SFD)

Experiment 1 (condition confusion simulation). To isolate the detection capability of SFD in a controlled setting, we replaced the degradation patterns of a target temperature in the NASA dataset with those of a different temperature. A normal generator adds small noise ($\sigma = 0.02$) to the real data, while a confused generator replaces data from a target temperature condition T_{fail} with patterns from a different temperature T_{wrong}. This setup mimics the scenario in which a generative model has failed to learn the degradation pattern at T_{fail} and instead outputs data resembling T_{wrong}. Table 6 reports the detection advantage (DA), defined as the ratio of the per-condition FD ratio to the FID ratio. DA quantifies how much more sensitively SFD detects a quality failure than FID does. The results reveal a clear inverse relationship between DA and minority proportion. For the 22°C condition, which comprises only 6.1% of the dataset, DA reaches 64–79, meaning that SFD's per-condition Fréchet distance is 64–79 times more sensitive than FID. For the near-majority 4°C condition (39.4%), the advantage is a modest 2.5 times. This pattern aligns with the dilution analysis of Section II: the rarer a condition, the more its failures are suppressed in FID, and the greater the advantage of per-condition evaluation. In practice, this alignment is consequential because the conditions most critical for safety are precisely those most likely to be underrepresented (Figure 2a).

Experiment 2 (CVAE generative model). Simulation establishes the principle. Real generative models, however, produce subtler quality differences that test whether SFD retains its advantage. Four CVAE variants were trained on the SNL dataset with progressively fewer temperature conditions (Table 7). The central finding concerns Model D, which was trained on 25°C data alone. FID registers a ratio of merely 1.01 times relative to the baseline Model A—a score that would give a practitioner no reason for concern. SFD reveals a different picture. The per-condition Fréchet distance at 15°C rises to 1.97 times and at 35°C to 1.84 times, while at 25°C (the sole training condition) it improves to 0.82 times. This profile indicates that Model D has overfit to 25°C at the expense of the excluded conditions.

Model B presents a scenario more likely to arise in practice: a researcher collects data predominantly at moderate temperatures and trains a model on what is available. FID's 1.02 times ratio would suggest that all is well. SFD's 1.25 times at the underrepresented 15°C condition, however, signals that further investigation is warranted. The ability

Table 7. CVAE results (SNL, 3-seed average). Ratios are relative to Model A.

Model	FID	SFD_c	FD(15°C)	FD(25°C)	FD(35°C)
A (all conditions)	12.52	22.47	23.52	15.06	10.06
Ratio relative to A					
B (15°C excl.)	1.02 times	1.09 times	**1.25 times**	0.97 times	0.98 times
C (label swap)	1.04 times	1.03 times	1.07 times	1.00 times	1.00 times
D (25°C only)	**1.01 times**	**1.43 times**	**1.97 times**	0.82 times	**1.84 times**

Table 8. Feature extractor comparison (SNL, CVAE Model D/A ratio).

Feature extractor	FID ratio	SFD_c ratio	FD(15°C)	FD(35°C)
Hand-crafted (17-dim.)	1.01 times	1.43 times	1.97 times	1.84 times
InceptionTime-style (32-dim.)	12.18 times	18.79 times	17.49 times	78.82 times
Autoencoder (16-dim.)	2.30 times	6.38 times	12.61 times	20.50 times

to flag graded, condition-specific degradation—rather than only catastrophic failure—is what makes SFD useful in real deployment (Figure 2b). This result provides empirical confirmation of the concern raised in the Introduction. A model that FID declares "no problem" is in fact generating inaccurate data under the minority conditions (15°C, 35°C) that are indispensable for safety evaluation. SFD_c resolves this blind spot by identifying *which* conditions are problematic.

Supplementary experiments on λ sensitivity (Experiment 3), cross-dataset generalization (Experiment 4), and the formal relationship between SFD and CFID (Experiment 5) are reported in Appendix C. Their principal conclusions are that $\lambda = 0$ maximizes per-condition sensitivity, that the detection advantage generalizes across all four datasets, and that SFD_c with $\lambda = 0$ reduces exactly to CFID. Inter-condition consistency is now assessed more directly by Layer 2 (CRC and CDR) with formal permutation tests. The Between-SFD term ($\lambda > 0$) is therefore no longer needed for this purpose, and $\lambda = 0$ is adopted as the recommended default.

Experiment 6 (feature extractor robustness). A potential objection is that SFD's advantage might depend on the choice of feature space. Table 8 addresses this concern by repeating the CVAE comparison with three different feature extractors. Under hand-crafted features, FID is 1.01 times while SFD reaches 1.97 times at 15°C. Under the InceptionTime-style CNN, FID rises to 12.18 times. SFD, however, reaches 78.82 times at 35°C. A better feature extractor raises the baseline sensitivity of both metrics. The diagnostic advantage of SFD—identifying *which* conditions fail—persists regardless of the feature space. A crucial point emerges from this comparison. Even with InceptionTime-style features that substantially boost FID's own sensitivity, FID cannot reveal *which* temperature condition is responsible for the quality gap. Improving the feature extractor raises the detection floor for FID, but the per-condition diagnostic information provided by SFD_c remains inaccessible to FID regardless of feature space sophistication. This confirms that SFD's advantage derives from its stratification structure, not from a particular feature design.

Experiment 7 (temporal-axis and condition$\times$time diagnosis). Experiments 1–6 demonstrated that condition-axis stratification (SFD_c) resolves FID's blind spot for minority-condition failures. However, as discussed in Section III, dilution also occurs along the temporal axis: if a generative model accurately reproduces the early portion of a degradation curve but fails in the late portion, SFD_c—which computes features over the entire curve—may miss this failure. Experiment 7 tests whether temporal stratification (SFD_t) and its combination with condition stratification ($SFD_{c\times t}$) can address this limitation. The

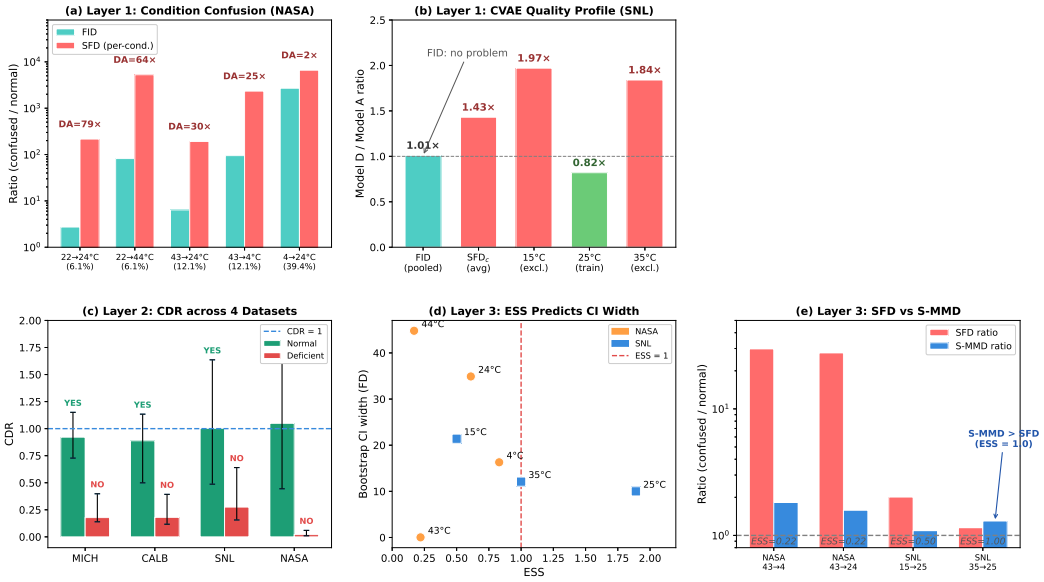

Figure 2. Overview of the three-layer diagnostic framework. **Top row** illustrates Layer 1 (quality diagnosis). (a) Experiment 1 (NASA condition confusion). SFD's per-condition Fréchet distance (red) detects condition-specific failures with up to 79 times higher sensitivity than FID (green). Detection advantage (DA) is inversely correlated with minority proportion. (b) Experiment 2 (SNL CVAE). FID registers 1.01 times (indistinguishable from baseline), while SFD reveals per-condition degradation at 15°C (1.97 times) and 35°C (1.84 times), with overfit improvement at 25°C (0.82 times). The top-right panel summarizes the three-layer workflow. **Bottom row** illustrates Layers 2 and 3. (c) CDR with bootstrap 95% confidence intervals across all four datasets. Normal generators (green) have CIs that include unity (YES), while deficient generators (red) exclude unity (NO), demonstrating that Layer 2 reliably separates correct from distorted inter-condition structure. (d) ESS predicts bootstrap CI width. Strata with ESS < 1 exhibit wide confidence intervals, confirming that ESS serves as a practical reliability indicator for Layer 1 estimates. (e) SFD and S-MMD comparison. At ESS $= 1.0$ (SNL 35→25), S-MMD ratio exceeds SFD ratio, confirming that S-MMD provides a more reliable signal when per-stratum samples are marginal. Figures 3 and 4 provide detailed three-layer evaluation for the HPC-FNO-CFM architecture (Experiment 9).

experiments above stratify by operating condition but treat each degradation curve as a whole. This is insufficient when the generative model reproduces the early phase of degradation correctly but fails in the late phase, where nonlinear acceleration beyond the knee point[26] is the dominant feature. SFD$_t$ addresses this by partitioning curves into temporal segments.

In a controlled simulation on NASA, we corrupted only the first or second half of the 43°C curves. SFD$_t$ with two segments assigns an elevated FD exclusively to the corrupted segment (173.7 times) while the intact segment registers precisely 1.00 times (Table 9). FID, which evaluates the full curve, cannot localize the failure to a specific phase.

Applying the same principle to the CVAE models yields the most physically informative result of the study. Table 10 presents the condition×time quality map for Model D relative to Model A. The largest degradation ratio, 8.69 times, appears in the latter half of the 35°C curves—precisely the regime where high-temperature degradation accelerates and a model trained only on 25°C data is least equipped to follow. The progression from FID (1.01 times) through SFD$_t$ (1.11 times) to SFD$_{c\times t}$ (3.18 times overall, 8.69 times at the critical cell) illustrates how increasing stratification granularity brings progressively more localized problems to light. For a practitioner, a single SFD$_{c\times t}$ table immediately reveals not only which temperature condition is problematic but also in which degradation phase the model fails.

Table 9. Partial confusion detection (NASA, 43°C→4°C, $K = 2$).

Corrupted segment	FID ratio	FD (corrupted)	FD (intact)
First half only	39.2 times	173.7 times	1.00 times
Second half only	79.3 times	70.7 times	1.00 times

Table 10. Condition×time quality map (SNL, CVAE Model D/A, $K = 2$).

Condition	First half	Second half
15°C (excluded)	4.45 times	3.79 times
25°C (trained)	0.75 times	0.91 times
35°C (excluded)	5.79 times	**8.69 times**

Experiment 8 (confounding detection). When temperature and C-rate are confounded in the experimental design, stratifying by temperature alone leaves C-rate effects entangled within each stratum. To test whether SFD can expose this confounding, we trained two CVAE models on the confound-free subset of SNL (1C and 2C at all three temperatures, 45 batteries): Model T conditioned on temperature alone and Model J conditioned on both temperature and C-rate. Evaluating both with $\text{SFD}_{\text{joint}}$ over six temperature×C-rate strata (Table 11), Model T exceeds Model J in every condition. To assess the stability of this result, the evaluation was repeated with $n_{\text{seeds}} = 10$ independent random initializations and a Mann-Whitney U test was applied to each condition. Model T exceeds Model J in all 10 seeds for every condition, with $p < 0.0001$ in all six conditions and a mean T/J ratio of 29.7 (SFD proxy; squared Euclidean distance between per-condition mean feature vectors). The largest gap occurs at 15°C/2C (T/J = 59.4), a condition whose combination of low temperature and high C-rate produces a degradation pattern most distinct from the 25°C average and therefore least well served by temperature-only conditioning. This comparison yields an actionable recommendation: temperature-only conditioning is insufficient for this dataset, and C-rate should be included as a conditioning variable. A second confounding source (cycle-count normalization) is analyzed in Experiment 8b (Appendix C).

Experiment 9 (alternative generative architecture: HPC-FNO-CFM). To verify that the framework's diagnostic capability extends beyond CVAE, we evaluated five variants of HPC-FNO-CFM[27], a conditional Flow Matching model that integrates a Fourier Neural Operator (FNO) pre-trained on battery physics. The five variants differ in how much of the pre-trained FNO is frozen during fine-tuning on NASA temperature extrapolation: `freeze_2layer` (two FNO layers frozen), `freeze_3layer` (three layers frozen), `freeze_1layer` (one layer frozen), `scratch_3layer` (no pre-training, trained from scratch with three constraint levels), and `pure_cfm` (no physical constraints, pure Flow Matching). Each variant generates voltage waveforms at five temperature conditions (4°C, 8°C, 13°C, 24°C, 43°C), where 24°C is the training condition and the others are extrapolation targets.

Table 12 reports the three-layer evaluation. SFD reveals a seven-order-of-magnitude range in per-condition quality: `freeze_2layer` achieves SFD = 2,353 with uniform quality across all five temperature conditions (FD ranges from 1,822 to 3,283), while `pure_cfm` reaches SFD = 1.8×10^{10} with catastrophic failure at 8°C (FD = 8.8×10^{10}). This enormous dynamic range demonstrates a key advantage of SFD: by evaluating each condition independently, it exposes failures that would be obscured in any pooled metric.

CRC provides a complementary perspective on inter-condition structure. `freeze_2layer` achieves CRC = 0.724, indicating that the ordering of inter-condition distances is substantially preserved. `pure_cfm`, by contrast, achieves CRC = 0.033, meaning that the physical structure across temperatures has been almost entirely destroyed. CDR reveals this destruction in detail: for `freeze_2layer`, 8 of 10 condition pairs have bootstrap confidence

Table 11. Confounding detection (SNL, 1C+2C, 45 batteries). Model T and Model J are each trained with $n_{\text{seeds}} = 10$ independent initializations. SFD proxy: squared Euclidean distance between per-condition mean feature vectors ($d = 5$ PCA components). T/J ratio: mean(SFD_T) / mean(SFD_J). Mann-Whitney U test (one-sided, H_1: T > J).

Condition	SFD_T mean	SFD_T std	SFD_J mean	SFD_J std	T/J	p-value
15°C/1C	0.01612	0.00247	0.00078	0.00140	20.8	< 0.0001◯
15°C/2C	0.00995	0.00147	0.00017	0.00008	59.4	< 0.0001◯
25°C/1C	0.00223	0.00034	0.00034	0.00037	6.5	< 0.0001◯
25°C/2C	0.01227	0.00099	0.00047	0.00028	25.9	< 0.0001◯
35°C/1C	0.00313	0.00056	0.00010	0.00009	30.8	< 0.0001◯
35°C/2C	0.00892	0.00128	0.00026	0.00019	34.8	< 0.0001◯
Mean					**29.7**	

Table 12. Three-layer evaluation of HPC-FNO-CFM variants on NASA temperature extrapolation. CDR "YES" indicates the number of condition pairs (out of 10) whose 95% bootstrap CI includes unity.

Model variant	SFD	CRC_R	CDR YES	S-MMD range
`freeze_2layer`	2,353	0.724	8/10	0.27–0.31
`freeze_3layer`	41,747	0.563	3/10	0.25–0.29
`freeze_1layer`	4.9M	0.306	0/10	0.24–0.28
`scratch_3layer`	186M	0.651	0/10	0.25–0.30
`pure_cfm`	17.7B	0.033	0/10	0.22–0.27

intervals that include unity (structure preserved), whereas for all other variants, every pair excludes unity (structure distorted). The only pairs where `freeze_2layer` fails are 43°C→8°C (CDR = 0.32) and 4°C→13°C (CDR = 0.26), which represent the most physically distant temperature combinations and are therefore the most challenging extrapolation targets.

A notable finding concerns S-MMD. Despite the seven-order-of-magnitude range in SFD, S-MMD values vary only between 0.22 and 0.31 across all five model variants. This confirms the theoretical prediction from Appendix A: MMD is insensitive to the covariance differences that dominate the Fréchet distance and cannot distinguish model quality in this setting. The practical implication is that S-MMD serves as a stable complement when sample sizes are small. SFD and CRC/CDR remain essential for capturing the quality differences that matter for temperature extrapolation.

This experiment confirms that the three-layer framework provides meaningful diagnostics for a generative architecture (conditional Flow Matching with FNO) that is fundamentally different from the CVAE used in Experiments 1–8. The ranking produced by SFD (`freeze_2layer` ≫ `freeze_3layer` ≫ `freeze_1layer` ≫ `scratch` ≫ `pure_cfm`) is consistent with the physical expectation that pre-trained FNO layers encode degradation physics that improves extrapolation, and that removing these constraints progressively degrades quality. Figure 3 visualizes the three-layer evaluation. Panel (a) shows the seven-order-of-magnitude SFD range on a logarithmic scale, making the progressive degradation from `freeze_2layer` to `pure_cfm` immediately visible. Panel (b) displays CRC values, where the drop from 0.724 to 0.033 quantifies the destruction of inter-condition structure. Panel (c) maps CDR for all 10 condition pairs across the five model variants, with green indicating preserved structure and red indicating distortion. Figure 4 complements this aggregate view by showing the per-condition Fréchet distance for each variant at each of the five temperatures. The uniformly low FD of `freeze_2layer` across all conditions

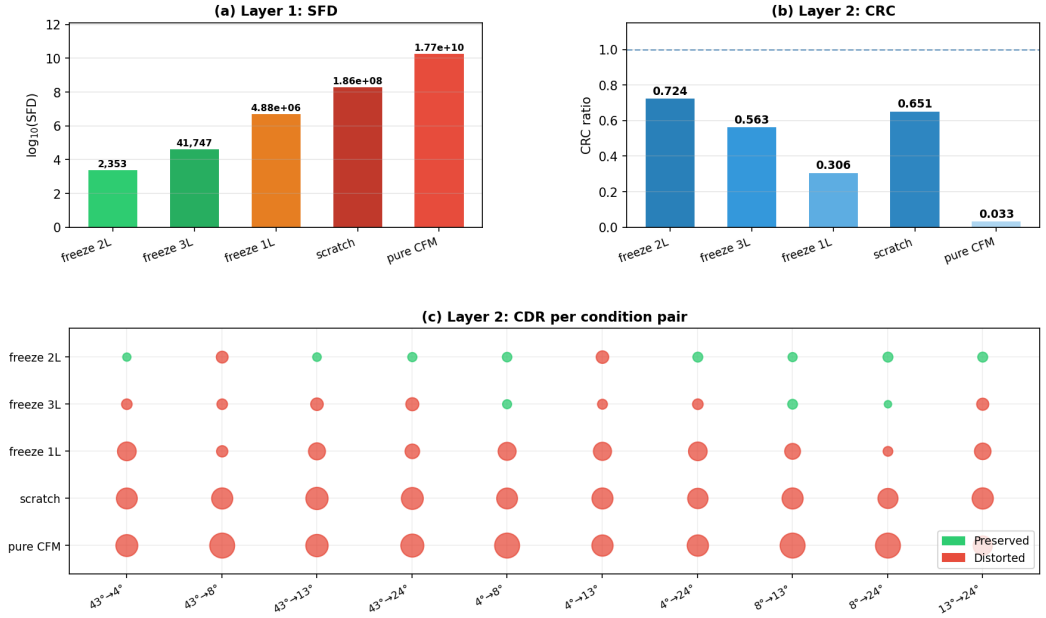

Figure 3. Three-layer evaluation of HPC-FNO-CFM model variants on NASA temperature extrapolation. (a) Layer 1 (SFD): the mean per-condition Fréchet distance spans seven orders of magnitude on a logarithmic scale, from 2,353 for `freeze_2layer` to 1.8×10^{10} for `pure_cfm`. Physical constraints from pre-trained FNO layers dramatically improve extrapolation quality. (b) Layer 2 (CRC): the inter-condition distance structure is well preserved by `freeze_2layer` (CRC = 0.724) but almost entirely destroyed by `pure_cfm` (CRC = 0.033). A dashed line marks perfect preservation (CRC = 1). (c) Layer 2 (CDR per condition pair): each circle represents one of 10 temperature-pair comparisons for a given model. Green circles indicate that the 95% bootstrap confidence interval for CDR includes unity (distance structure preserved); red circles indicate exclusion (structure distorted). Bubble size reflects the magnitude of $|\log_{10}(\text{CDR})|$. Only `freeze_2layer` preserves the majority of condition pairs (8 of 10); all other variants distort every pair or nearly every pair.

contrasts with the extreme condition-dependent variation exhibited by the other variants, most notably the catastrophic failure of `pure_cfm` at 8°C.

5.2. Layer 2: Inter-Condition Structure Verification (CRC and CDR)

Layer 1 identifies *where* generation quality breaks down but does not assess whether the generative model preserves the physical relationships between conditions. A model might produce plausible curves at each temperature yet distort the relative distances—for example, making 43°C and 25°C degradation patterns more similar than they are in reality. CRC and CDR address this complementary question.

Table 13 reports CRC values computed in parallel with the Layer 1 experiments using mean-vector Euclidean distance and 200 generated samples per condition. Normal generators consistently achieve CRC above 0.87, indicating that the ordering of inter-condition distances is well preserved. Deficient generators (confused or single-condition models) produce CRC values of 0.53–0.73, reflecting measurable distortion of the physical structure. When cross-stratification increases the number of condition pairs, permutation tests reach strong significance: $p = 0.004$ for NASA and $p = 0.012$ for SNL in Experiment 7, and $p = 0.001$ for the temperature×C-rate stratification in Experiment 8. The NASA result merits emphasis. Despite per-condition sample sizes of only $n = 3$–15, which render SFD estimates unreliable, CRC with mean-vector distance and temperature×time cross-stratification achieves $p = 0.004$. This illustrates a key practical benefit of the three-layer design: when Layer 1 estimates are uncertain due to small samples, Layer 2 can still provide statistically grounded conclusions about structural integrity.

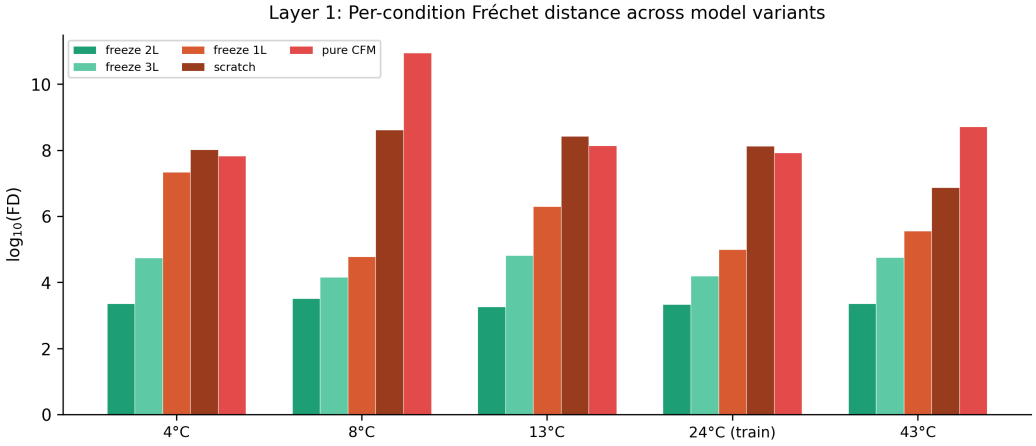

Figure 4. Per-condition Fréchet distance for each HPC-FNO-CFM variant (log scale). `freeze_2layer` achieves uniformly low FD across all five temperature conditions (range: 1,822–3,283), while other variants exhibit large condition-dependent variation. The `pure_cfm` model fails catastrophically at 8°C (FD $= 8.8 \times 10^{10}$), a failure invisible to any pooled metric but immediately apparent in the per-condition SFD profile.

Table 13. CRC results (mean-vector distance, $n_{\text{gen}} = 200$).

Experiment	Generator	CRC_R	p-value
Exp 1 (NASA)	Normal	0.946	—
	43→24 confused	0.727	—
	43→4 confused	0.656	—
Exp 2 (SNL)	Model A (all)	0.902	—
	Model B (15°C excl.)	0.531	—
	Model D (25°C only)	0.654	—
Exp 7 (SNL $c \times t$)	Normal	0.870	0.012
	Model D	0.611	0.311
Exp 7 (NASA $c \times t$)	Normal	0.870	0.004
	Confused	−0.281	0.802
Exp 8 (SNL)	Temp-only ($K=3$)	0.949	0.067
	Joint ($K=6$)	0.927	0.001

Table 14 extends the analysis to CDR, which applies even when only two conditions are available. Across all four datasets, normal generators yield CDR values close to unity with 95% bootstrap confidence intervals that include one, while deficient generators produce CDR values of 0.02–0.28 with intervals that exclude one. The MICH dataset ($K = 2$, 25°C and 45°C only) is the most important test case: CRC cannot be computed for two conditions, yet CDR reliably distinguishes the normal generator (CDR = 0.921, CI includes 1) from the confused generator (CDR = 0.178, CI excludes 1). When temporal cross-stratification is applied to MICH ($K = 4$, 6 pairs), the resulting CRC-ratio separates the two generators further (0.952 vs. 0.395), confirming the structural distortion through multiple independent pairs. For a practitioner working with a two-condition dataset, CDR thus provides a single, interpretable number with a confidence interval that answers the question: "Does my generative model preserve the distance between these two conditions?" (Figure 2c).

Experiment 10 (MOP: condition-ordering verification). Experiments 1–9 validate Layer 1 (SFD) and the distance-preservation aspects of Layer 2 (CRC, CDR). Experiment 10 evaluates MOP, which addresses a complementary question about whether the generative model preserves the *ordering* of conditions along the conditioning axis. This property

Table 14. CDR results (mean-vector distance, $n_{\text{gen}} = 200$).

Dataset	Generator	CDR	CI_{lo}	CI_{hi}	1 in CI
MICH	Normal	0.921	0.728	1.151	YES
	Confused 25→45	0.178	0.139	0.398	NO
CALB	Normal (0→45)	0.888	0.499	1.134	YES
	Confused 0→45	0.180	0.117	0.393	NO
SNL	Normal (15→35)	1.003	0.487	1.636	YES
	Model D (15→35)	0.275	0.156	0.640	NO
NASA	Normal (4→43)	1.049	0.444	1.725	YES
	Model D (4→43)	0.021	0.010	0.060	NO

cannot be assessed by CRC when the real data distance matrix is non-monotone, a common situation in battery datasets where cell-to-cell variability can obscure temperature effects.

Ten independent CVAE models were trained on the SNL 1C+2C dataset ($K = 6$ conditions, $d = 5$ PCA features). MOP was computed for four generator types, namely correctly conditioned (normal), temperature-label-reversed (type-2 inversion), C-rate-label-permuted (type-3 confusion), and condition-free (type-4 uniform). A one-sided Mann-Whitney U test compared the MOP distribution of each defective generator against the normal reference.

The normal generator achieves a stable MOP of $+0.971 \pm 0.029$ across all 10 seeds, confirming that the SNL temperature axis is reliably captured by the projection direction $\mathbf{v}_1$ ($\rho_{\text{proj}} = 0.717$). Detection is 100% for all three defect types ($p < 0.0001$). The type-2 inversion generator produces MOP $= -0.400$, which falls below the threshold of zero and is therefore flagged as inverted. The type-3 confusion generator produces MOP $= +0.406$, which falls below the dynamic threshold $\mu_{\text{normal}} - 0.5 = 0.471$ and is flagged as confused. The type-4 uniform generator produces MOP $= +0.109$ with substantially higher variance (std $= 0.330$) than the other types, because a condition-free generator occasionally reproduces partial temperature ordering by chance. The Mann-Whitney test nevertheless achieves $p < 0.0001$ because the distribution of MOP values is systematically shifted below the normal reference.

The cross-dataset generalization of MOP (four datasets), the minimum K required for CRC significance, and the root cause of CDR inflation are examined in Experiments 11–14 (Appendix C).

5.3. Layer 3: Estimation Reliability (S-MMD and ESS)

The final layer addresses the question that underlies all per-stratum evaluation: when sample sizes are small, can the Fréchet distance estimates from Layer 1 be trusted? Table 15 compares SFD and S-MMD ratios for the condition confusion experiments. Both metrics agree on the direction of quality degradation in every case, providing mutual corroboration. Their magnitudes, however, diverge markedly: SFD ratios reach 29.8 times while S-MMD ratios remain below 2 times. This divergence is consistent with the theoretical analysis of Appendix A. SFD captures covariance mismatch, which produces large ratios when genuine distributional differences exist. It also amplifies estimation noise when per-stratum samples are few. S-MMD, by contrast, does not estimate covariance matrices and therefore remains stable even at small n, at the cost of reduced sensitivity to covariance structure.

The complementary value of S-MMD is most evident in the SNL 35→25 confusion. The 35°C stratum has ESS $= 1.0$, placing it at the margin of reliable Fréchet distance estimation. Here S-$\text{MMD}_r = 1.30$ exceeds $\text{SFD}_r = 1.15$, indicating that MMD detects a quality difference that the noisy FD estimate underestimates (Figure 2e). In practice, when

Table 15. SFD and S-MMD comparison (PCA $d = 5$).

Dataset / Pattern	SFD_r	$S\text{-}MMD_r$	ConsI	ConsI CI
NASA 43→4	29.8	1.82	0.40	[−0.31, 1.0]
NASA 43→24	27.7	1.58	0.40	[−0.31, 1.0]
SNL 15→25	2.01	1.09	0.50	[0.50, 1.0]
SNL 35→25	1.15	1.30	1.00	[−0.50, 1.0]

Table 16. Bootstrap 95% CIs for SFD and CRC. CIs narrow as ESS increases.

Dataset	Generator	SFD [CI]	CRC_R [CI]
NASA	Normal	7.6 [1.0, 17.4]	0.33 [−0.61, 0.91]
	Confused	14.4 [1.5, 28.2]	−0.02 [−0.80, 0.54]
	Model D	20.0 [3.7, 38.4]	0.14 [−0.78, 0.77]
SNL	Normal	17.5 [13.7, 22.8]	0.83 [0.56, 0.97]
	Confused	19.8 [16.0, 24.7]	0.73 [0.50, 0.96]
	Model D	23.5 [19.7, 29.9]	0.70 [0.32, 0.95]

a practitioner observes a low ESS for a particular stratum (Figure 2d), S-MMD should be consulted alongside SFD, and discrepancies between the two should be resolved in favor of the more stable estimate.

To quantify the estimation uncertainty of SFD itself, we computed bootstrap 95% confidence intervals by resampling real and generated data within each stratum (500 iterations). Table 16 reports the results for three generators on both datasets. On the NASA dataset, where all strata have ESS < 1, the SFD confidence intervals are wide: the normal generator yields SFD $= 7.6$ with a CI of $[1.0, 17.4]$. On the SNL dataset, where ESS ranges from 0.50 to 1.89, the intervals are considerably tighter: Normal SFD $= 17.5$ $[13.7, 22.8]$. Despite the wide CIs on NASA, the ordering Normal $<$ Confused $<$ Model D is preserved in the point estimates across both datasets, indicating that the relative ranking of generators is robust even when absolute values are uncertain.

The CI width correlates strongly with ESS. For the NASA 44°C stratum (ESS $= 0.17$, $n = 3$), the per-condition FD CI spans $[0.8, 45.7]$—a range so wide that the point estimate alone is uninformative. For the SNL 25°C stratum (ESS $= 1.89$, $n = 34$), the CI is $[9.4, 19.4]$, narrow enough for practical use. This confirms that ESS serves as a reliable predictor of when SFD estimates can be trusted.

When Layer 1 CIs are wide, Layer 2 metrics provide sharper discrimination. On NASA with the Model D generator, the SFD CI of $[3.7, 38.4]$ overlaps substantially with the Normal CI of $[1.0, 17.4]$, leaving the quality difference statistically ambiguous. CDR, however, excludes unity for all six condition pairs (e.g., $CDR(4 \to 43) = 0.029$ $[0.011, 0.064]$), providing unambiguous evidence that Model D has distorted the inter-condition structure. This complementarity between layers—Layer 1 for diagnostic localization, Layer 2 for statistical confirmation—is the practical payoff of the three-layer design.

6. Discussion

6.1. The Role of Data Scarcity in Evaluation Design

The experiments reported above consistently show that the three-layer framework detects quality problems that FID cannot. The most important factor underlying this advantage is not stratification per se—CFID already stratifies by condition—but the combination of stratification with tools specifically designed for the data-scarce regime. Battery degradation datasets contain tens of cells, not tens of thousands. Per-condition strata may contain as few as three or four batteries. In this regime, covariance-based metrics such

Table 17. Detection performance by stratification granularity (SNL, CVAE Model D/A ratio).

Stratification	Diagnostic question	D/A ratio
None (FID)	What is the overall quality?	1.01 times
Condition c (SFD_c)	Which conditions fail?	1.97 times
Time k (SFD_t)	Which time window fails?	1.11 times
$c \times k$ ($SFD_{c\times t}$)	Which condition in which window?	3.18 times
$c_1 \times c_2$ (CI)	Is there inter-condition confounding?	1.72 times

as the Fréchet distance become unreliable, and the practitioner faces a dilemma: evaluate per-condition quality with an unreliable metric, or aggregate across conditions and lose diagnostic resolution.

The three-layer framework resolves this dilemma by separating the diagnostic question (Layer 1, what is the quality?), the structural question (Layer 2, is the physical structure preserved?), and the reliability question (Layer 3, can the estimate be trusted?). When Layer 1 estimates are uncertain because ESS is low, Layer 2 can still provide statistically grounded conclusions through CRC or CDR with mean-vector distance and permutation tests. When both layers are uncertain, Layer 3 flags the problem through low ESS and ConsI. This separation of concerns is what makes the framework useful in practice: no single metric is required to answer all three questions simultaneously.

6.2. Stratification Granularity and Its Limits

Table 17 summarizes the detection performance of each SFD variant in the CVAE Model D experiment.

The progression from 1.01 times to 3.18 times illustrates a general principle: finer stratification reveals more localized quality problems. Starting from a situation where FID judges the model as "no problem" at 1.01 times, stratification by condition reveals a 1.97 times quality gap, and adding the temporal axis raises it further to 3.18 times. This progressive increase demonstrates that the framework delivers diverse diagnostics through a single control parameter: the choice of stratification variable. However, finer stratification also reduces the number of samples per stratum, eventually degrading estimation reliability. In our experiments, temporal segmentation beyond $K = 3$ produced unstable Fréchet distance estimates. The practitioner must therefore choose stratification granularity to align with physically meaningful divisions—temperature conditions, early versus late degradation phases, charge rates—while ensuring that each stratum retains enough samples for reliable estimation. ESS provides a quantitative guide for this decision.

A natural question is whether temporal segments should be defined by the knee point—the physically motivated boundary between Phase 1 (linear SEI growth) and Phase 2 (nonlinear acceleration)—rather than by equal intervals. We compared four strategies on the NASA and SNL datasets: equal-interval $K = 2$ and $K = 3$, and knee-point-adaptive $K = 2$ and $K = 3$ (Experiment 7b, Appendix C). Equal-interval $K = 2$ achieved a confused-to-normal SFD ratio of 153.7 times on NASA, compared to 14.7 times for knee-point $K = 2$. The adaptive strategy yielded a lower baseline SFD under normal generation (0.90 vs. 5.39), suggesting fewer false alarms. Its detection power, however, was substantially lower.

The cause is the large inter-battery variability in knee-point position. In the NASA 4°C condition, knee-point positions range from 14% to 82% of the degradation curve (standard deviation 0.26). When each battery is split at its own knee point, the "pre-knee" segment for one battery contains 14% of the curve while for another it contains 82%. Data from physically distinct degradation stages are pooled under the same label, destabilizing the Fréchet distance estimate. By contrast, equal-interval splitting applies the same boundary

to all batteries, preserving sample comparability within each segment even at the cost of physical precision.

This finding reflects a broader tension in the data-scarce regime: physical fidelity in segmentation must be balanced against statistical stability. With large datasets (e.g., the 124 cells of Severson et al.[4]), grouping batteries by similar knee-point position before applying adaptive splitting could recover the physical advantage. With the sample sizes typical of current battery degradation datasets, however, equal-interval splitting provides a more reliable basis for evaluation, and we adopt it as the default.

6.3. Novelty of CRC, CDR, and MOP

CRC and CDR share the motivation of Benny et al.'s between-class evaluation [28], which compares the distribution of class-average vectors between real and generated data. CRC and CDR differ in that they specifically test whether the *rank ordering* (CRC) and *ratio* (CDR) of inter-condition distances are preserved—a property that between-class FID does not directly measure—and apply permutation testing to provide statistical significance in the small-K regime typical of battery datasets. MOP further asks whether the model preserves the *ordering* of conditions along the conditioning axis, which CRC cannot answer when the real data distance matrix is non-monotone.

The three Layer-2 indicators are complementary by design. CRC detects type-3 confusion (scrambled distance ordering, requires $K \geq 8$). CDR detects type-4 uniform collapse (CDR $\ll 1$, requires $n \geq 10$ per condition). MOP detects type-2 inversion and type-3 confusion, and remains effective when CRC lacks statistical power due to small K. Cross-validated on the SNL, NASA, CALB, and MICH battery datasets with $n_{\text{seeds}} = 10$, MOP achieved 100% detection of type-2 and type-3 defects on three of four datasets ($p < 0.001$). CDR achieved 100% detection of type-4 defects on datasets with sufficient samples. The two metrics showed no cross-sensitivity, confirming orthogonal coverage.

A theoretical lower bound on the CRC permutation test shows that for K conditions and $M = \binom{K}{2}$ pairs, $p_{\min} = 1/M!$, giving $p_{\min} = 0.167 > 0.05$ at $K = 3$ and $p_{\min} < 10^{-3}$ at $K \geq 5$. This explains empirically why CRC fails for small-K datasets and motivates MOP as the primary detection mechanism in such cases.

6.4. Positioning SFD Within the Landscape of Existing Metrics

The mathematical equivalence between SFD_c ($\lambda = 0$) and CFID means that SFD does not supersede prior work but rather extends it. CFID provides detection power through condition-axis stratification; SFD generalizes this with temporal stratification (SFD_t), condition$\times$time stratification ($\text{SFD}_{c\times t}$), and confounding detection (CI). In earlier formulations, the Between-SFD term ($\lambda > 0$) additionally served to assess inter-condition consistency; in the present three-layer framework, this role is handled more directly and with formal statistical significance by Layer 2 (CRC and CDR), and $\lambda = 0$ is adopted as the default. All of these capabilities are realized by varying the stratification variable s in Eq. (3), requiring no new mathematical apparatus beyond the core SFD formulation.

The feature extractor comparison (Experiment 6) confirmed that SFD's advantage is independent of the feature space. With InceptionTime-style features, FID's own detection power reaches 12.18 times, yet SFD_c's per-condition FD surpasses it further still (FD(35°C) = 78.82 times). This result provides experimental support for the analysis in Section II: FID's limitation originates not in the inadequacy of the feature extractor but in the manner of aggregation—collapsing a mixture distribution into a single scalar.

CRC and CDR are motivated by Benny et al.'s between-class evaluation [28], which compares the distribution of class-average vectors between real and generated data. CRC and CDR extend this direction by specifically testing whether the *rank ordering* (CRC) and

magnitude ratio (CDR) of inter-condition distances are preserved. FID, CFID, FJD, and AMMD all evaluate quality within or across conditions but do not assess whether the relative distances between conditions—which encode physical relationships such as the monotonic acceleration of degradation with temperature—are faithfully reproduced in rank order or magnitude.

The practical importance of this question is evident in the CDR results for MICH. With only two temperature conditions, per-condition SFD can indicate whether quality is adequate at each temperature. It cannot, however, reveal whether the model compresses or exaggerates the distance between 25°C and 45°C. CDR fills this gap with a single interpretable number and a confidence interval. For datasets with three or more conditions, CRC extends this assessment to the full pairwise distance matrix, and cross-stratification provides the statistical power needed for formal significance testing.

6.5. Variance Decomposition and the Source of Dilution

The mathematical structure of dilution can be understood through the classical covariance decomposition. The mixture covariance Σ_{mix} decomposes into a within-group component $\Sigma_W = \sum_c p(c)\Sigma_c$ and a between-group component $\Sigma_B = \sum_c p(c)(\mu_c - \bar{\mu})(\mu_c - \bar{\mu})^T$. FID operates on $\Sigma_{\text{mix}} = \Sigma_W + \Sigma_B$, so changes in a particular condition's covariance Σ_c can be dwarfed by the magnitude of Σ_B. SFD computes the Fréchet distance directly from each Σ_c, bypassing Σ_B entirely. The mutual information $I(X;C) = H(X) - H(X|C)$ quantifies how strongly the condition parameter influences the distribution. The larger $I(X;C)$, the more severe the dilution in FID and the greater the benefit of stratified evaluation. Battery degradation data, where degradation patterns differ qualitatively across temperatures, represent a domain with high $I(X;C)$.

This connection to entropy can be made precise. The entropy of the mixture distribution satisfies $H(X) \geq H(X|C)$, with the gap equal to the mutual information $I(X;C)$. FID effectively measures a whole-distribution distance corresponding to $H(X)$. The larger $I(X;C)$ is, the more conditions contribute to the mixture and the coarser this evaluation becomes. Within-SFD instead measures a per-condition distance corresponding to $H(X|C)$ and is unaffected by $I(X;C)$. This structural analogy yields an important practical guideline: **the larger $I(X;C)$ is for a given dataset, the more severely FID is affected by dilution and the greater the benefit of stratified evaluation**.

6.6. Applicability to Other Data-Scarce Domains

The framework is designed for domains that share three characteristics with battery degradation: scarce data, condition-dependent qualitative changes, and safety-critical minority conditions. We identify three such domains where the framework can be applied with minimal modification.

In **materials science**, stress–strain curves and thermal analysis profiles vary with chemical composition and processing parameters (temperature, pressure, atmosphere). Data for extreme compositions or high-temperature processes are costly to acquire and inherently scarce, creating the same minority-condition dilution structure observed in battery data. SFD_c stratified by processing condition, combined with CDR for two-condition datasets, would enable per-condition quality assessment of synthetic materials data.

In **structural health monitoring**, sensor time series from bridges, turbines, and aircraft components are recorded under varying load and environmental conditions. Data under extreme loads or rare fault conditions are sparse, and generative models used for data augmentation must reproduce condition-specific failure signatures. SFD_t and $\text{SFD}_{c\times t}$ are directly applicable to temporally localized quality assessment of synthetic fault data.

In **nuclear engineering**, reactor sensor data under transient and accident scenarios are extremely scarce—often limited to a handful of simulation runs per scenario. Generative models trained on normal-operation data and asked to extrapolate to accident conditions face the same evaluation challenge as battery models extrapolating to extreme temperatures. The combination of CDR (for two-scenario comparisons) and ESS (for flagging unreliable strata) would provide quality assurance in this safety-critical setting.

These domains share the property that data scarcity is not a temporary inconvenience but a permanent structural feature of the problem, making evaluation tools designed for large datasets fundamentally inadequate.

6.7. Extreme Temperature Considerations

The NASA dataset includes batteries tested at 4°C, a temperature at which lithium plating on the anode becomes the dominant degradation mechanism rather than SEI growth[2]. This qualitative change in degradation physics means that 4°C curves are not merely attenuated versions of 24°C curves but follow a fundamentally different trajectory. SFD_c captures this distinction because it evaluates each temperature condition independently. A practitioner should be aware, however, that the physical meaning of a large FD at 4°C differs from that at 43°C: the former reflects failure to reproduce plating-dominated degradation, while the latter reflects failure to reproduce thermally accelerated SEI decomposition. CRC and CDR complement this interpretation by assessing whether the generative model preserves the distance between these qualitatively distinct regimes. If CDR for the 4°C–43°C pair deviates substantially from unity, the model has failed to capture the most safety-relevant contrast in the dataset. Extending the framework to sub-zero temperatures, where additional mechanisms such as electrolyte freezing and lithium dendrite formation become relevant, is an important direction for future work in cold-climate battery applications.

6.8. Limitations

The framework assumes discrete strata. When the conditioning variable is continuous (e.g., temperature in 0.1°C increments), binning is required, and the choice of bin width influences the results. Extending the framework to continuous stratification via kernel density estimation is an important direction for future work.

Our experimental validation primarily uses CVAE models for the controlled condition-exclusion experiments (Experiments 1–8). Although Experiment 9 evaluates an HPC-FNO-CFM model (conditional Flow Matching with Fourier Neural Operator), confirming that the framework generalizes beyond CVAE, further validation with additional architectures such as TimeGAN, TimeVAE, and diffusion-based models would strengthen the generality claims. We emphasize, however, that SFD is a property of the *evaluation metric*, not of the generative model. The dilution problem analyzed in Section II arises from FID's aggregation structure and is independent of how the data were generated. Thus, we expect the framework's advantages to persist across generative architectures, though experimental confirmation with other model families is a natural direction for future work.

The CRC and CDR validation in the CVAE experiments (Experiments 1–8) relies on simulated generators (normal, confused, distorted) rather than on trained generative models with genuine structural deficiencies. Experiment 9 applies CRC and CDR to a real trained model (HPC-FNO-CFM), providing a partial remedy; however, further validation with trained CVAE, GAN, and diffusion-based models exhibiting subtle structural distortions would strengthen the claims.

Seed dependence. A natural concern is whether the MOP and CDR results depend on a particular random initialization of the generative model. To address this, the MOP experiments were conducted with $n_{\text{seeds}} = 10$ independent random seeds (Experiment 11).

Across all 10 seeds, the detection rate for type-2 inversion remained 100% (Mann-Whitney $p < 0.0001$) and the reference model's MOP was stable at $+0.971 \pm 0.029$. The CDR experiments were repeated with $n_{\text{seeds}} = 5$. The detection rate for type-4 uniform collapse was 100% on all datasets with sufficient per-condition samples ($n \geq 10$). These results confirm that the diagnostic conclusions are not artifacts of a particular random seed.

When PCA is used to reduce the feature dimension for S-MMD (as recommended for strata with low ESS), the projection is fitted on the pooled data and then applied to each stratum. This introduces a potential bias: the principal components are dominated by the majority stratum, and minority strata may be poorly represented in the projected space. In our experiments with $d = 5$ PCA components, the first five components explain 85–92% of the total variance but may capture less of the variance specific to the 43°C stratum ($n = 4$). When the Fréchet distance is computed in a regularized, low-dimensional space, the regularization parameter ϵ can dominate the estimate for extremely small strata, producing artificially stable values that mask genuine instability (as observed in the 43°C FD_CV $= 0.12$ anomaly in the theoretical verification experiments). Practitioners should interpret FD values for strata with ESS < 1 as order-of-magnitude indicators rather than precise measurements, and should rely on S-MMD and CDR for confirmatory evidence.

Finally, the theoretical analysis in Appendix A assumes Gaussian distributions and bounded kernels. Battery degradation data may violate these assumptions, particularly in the tails of the distribution where safety-critical degradation patterns reside.

The choice of feature extractor also substantially affects SFD values. As Experiment 6 demonstrated, the FID ratio increased from 1.01 times with hand-crafted features (17-dim.) to 12.18 times with InceptionTime-style features (32-dim.)—a change of over an order of magnitude depending on the feature space. While the structural advantage of SFD (per-condition FD $>$ FID) holds regardless of the feature extractor, the absolute detection sensitivity is strongly feature-dependent. Leveraging large-scale pre-trained time series encoders, or developing domain-agnostic feature extractors, are important avenues for improving the framework's generality.

The Layer-2 indicators (CRC, CDR, and MOP) each carry applicability conditions that must be verified before the results can be interpreted reliably. In brief, CDR requires at least $n = 10$ cells per condition; at smaller sample sizes, the inter-condition distance estimate is dominated by sampling noise and CDR is inflated regardless of model quality. CRC requires at least $K = 8$ conditions for the permutation test to achieve conventional significance, and additionally requires the real data distance matrix to be monotone with respect to the condition ordering. MOP requires a projection quality of $\rho_{\text{proj}} \geq 0.7$, which measures how well the temperature axis is captured in the feature space; when cell-to-cell variability overwhelms the temperature signal, MOP sensitivity degrades. Crucially, these are constraints on the evaluation data, not on the generative model: a negative result from an indicator may reflect insufficient data rather than a well-performing generator. Appendix D provides a detailed treatment of these applicability conditions, including empirically derived thresholds, the failure modes identified through controlled experiments, and a worked example showing what each indicator could and could not determine on the NASA battery dataset.

7. Conclusion

We have proposed a three-layer diagnostic framework centered on **Stratified Fréchet Distance (SFD)** for resolving the dilution problem of FID in the evaluation of conditional time series generation models under data scarcity. FID aggregates all data into a single score, causing it to overlook quality degradation in safety-critical minority conditions and in late-

cycle regions where degradation accelerates. The framework addresses this fundamental limitation by decomposing the evaluation into three complementary questions.

Layer 1 (SFD) diagnoses *where* and *when* quality breaks down. In a setting where FID registers virtually no difference from the baseline (ratio 1.01), SFD_c detects per-condition degradation 1.97 times larger than the baseline, and $\text{SFD}_{c\times t}$ localizes the largest quality gap (8.69 times the baseline) to the latter half of 35°C curves—precisely the regime of accelerated high-temperature degradation most relevant to safety. This result calls for a reconsideration of the current practice of relying on FID as the sole evaluation metric, particularly from the standpoint of safety. On a conditional Flow Matching model with physics-informed constraints, SFD spans seven orders of magnitude across model variants, correctly ranking them by physical constraint level.

Layer 2 (CRC, CDR, and MOP) verifies *whether the physical structure across conditions is preserved*. CRC achieves $p = 0.004$ on the NASA dataset despite per-condition sample sizes of only 3–15, demonstrating that structural verification is possible even when per-condition Fréchet distance estimates are unreliable. MOP detects condition-ordering defects that CRC cannot identify when K is small or the real data distance matrix is non-monotone. Across four battery datasets with $n_{\text{seeds}} = 10$, MOP achieves 100% detection of type-2 inversion and type-3 confusion defects on three of four datasets ($p < 0.0001$). CDR extends this capability to two-condition datasets, distinguishing normal generation (CDR = 0.92) from structurally deficient generation (CDR = 0.18) with confidence intervals that exclude unity. The Confounding Index further reveals that temperature–C-rate confounding degrades generation quality by a factor of 1.72 times, providing actionable guidance for the model design decision of which conditioning variables to include.

Layer 3 (ESS/S-MMD) assesses *how reliable each per-stratum estimate is*. S-MMD detects quality differences that SFD misses when ESS is low, confirming its complementary role. Bootstrap confidence intervals confirm that ESS reliably predicts when Layer 1 estimates can be trusted, and that Layer 2 metrics provide statistically grounded conclusions even when Layer 1 estimates are uncertain.

These findings have been confirmed through fourteen experiments across four battery datasets (161 cells), three feature extractors, and two generative architectures (CVAE and conditional Flow Matching with Fourier Neural Operator). Its applicability extends to other data-scarce, physics-governed domains—materials science, structural health monitoring, nuclear engineering—where scarce data, condition-dependent qualitative changes, and safety-critical minority conditions create the same evaluation challenge. The framework raises a fundamental question that has been largely overlooked. Satisfactory average quality as measured by FID does *not* guarantee adequate quality under every condition and in every temporal region. Taken together, these findings establish that average quality as measured by FID provides no guarantee of per-condition adequacy. In safety-critical applications, per-condition adequacy is precisely what is needed. The three-layer framework provides the means to verify it through stratification, structural verification, and reliability assessment.

References

1. Roman, D.; Saxena, S.; Robu, V.; Pecht, M.; Flynn, D. Machine learning pipeline for battery state-of-health estimation. *Nature Machine Intelligence* **2021**, *3*, 447–456.
2. Plett, G.L. *Battery Management Systems, Volume II: Equivalent-Circuit Methods*; Artech House, 2015.
3. Han, X.; Lu, L.; Zheng, Y.; et al. A review on the key issues of the lithium ion battery degradation among the whole life cycle. *eTransportation* **2019**, *1*, 100005.
4. Severson, K.A.; Attia, P.M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M.H.; Aber, M.; Chueh, W.C.; Ermon, S.; et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy* **2019**, *4*, 383–391.

5. Dos Reis, G.; Strange, C.; Sheridan, M.; Hynes, V. Lithium-ion battery data and where to find it. *Energy and AI* **2022**, *5*, 100081.
6. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, 2014, Vol. 27.
7. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* **2014**.
8. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the Advances in Neural Information Processing Systems, 2020, Vol. 33, pp. 6840–6851.
9. Lipman, Y.; Chen, R.T.Q.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow matching for generative modeling. In Proceedings of the International Conference on Learning Representations, 2023.
10. Brophy, E.; Wang, Z.; She, Q.; Ward, T. Generative adversarial networks in time series: A systematic literature review. *ACM Computing Surveys* **2023**, *55*, 1–31.
11. Yoon, J.; Jarrett, D.; van der Schaar, M. Time-series generative adversarial networks. In Proceedings of the Advances in Neural Information Processing Systems, 2019, Vol. 32.
12. Desai, A.; Freeman, C.; Wang, Z.; Beaver, I. TimeVAE: A variational auto-encoder for multivariate time series generation. In Proceedings of the arXiv preprint arXiv:2111.08095, 2021.
13. Tashiro, Y.; Song, J.; Song, Y.; Ermon, S. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In Proceedings of the Advances in Neural Information Processing Systems, 2021, Vol. 34.
14. Eivazi, H.; Hebenbrock, A.; Wildermuth, R.; et al. DiffBatt: A diffusion model for battery degradation prediction and synthesis. *arXiv preprint arXiv:2410.23893* **2024**.
15. Wu, B.; Widanage, W.D.; Yang, S.; Liu, X. Battery digital twins: Perspectives on the fusion of models, data and artificial intelligence for smart battery management systems. *Energy and AI* **2020**, *1*, 100016.
16. Howey, D.A.; Roberts, S.A.; Viswanathan, V.; et al. Enabling battery digital twins at the industrial scale. *Joule* **2023**, *7*, 928–934.
17. Pang, G.; Shen, C.; Cao, L.; van den Hengel, A. Deep learning for anomaly detection: A review. *ACM Computing Surveys* **2021**, *54*, 1–38.
18. Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S.P. A critical review of machine learning of energy materials. *Advanced Energy Materials* **2019**, *10*, 1903242.
19. Mayer, P.; Luzi, L.; Siahkoohi, A.; Johnson, D.H.; Baraniuk, R.G. Improving fairness and mitigating MADness in generative models. *arXiv preprint arXiv:2405.13977* **2024**.
20. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30.
21. Stein, G.; Cresswell, J.C.; Hosseinzadeh, R.; Sui, Y.; Ross, B.L.; Villecroze, V.; Liu, Z.; Caterini, A.L.; Taylor, J.E.T.; Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems* **2023**, *36*.
22. Soloveitchik, M.; Diskin, T.; Morin, E.; Wiesel, A. Conditional Fréchet inception distance. *arXiv preprint arXiv:2103.11521* **2021**.
23. DeVries, T.; Romero, A.; Pineda, L.; Taylor, G.W.; Drozdzal, M. On the evaluation of conditional image generation. In Proceedings of the arXiv preprint arXiv:1907.08175, 2019.
24. Saha, B.; Goebel, K. Battery data set. Technical report, NASA Ames Prognostics Data Repository, 2007.
25. Tan, R.; Hong, W.; Tang, J.; Lu, X.; Ma, R.; Zheng, X.; Li, J.; Huang, J.; Zhang, T.Y. BatteryLife: A comprehensive dataset and benchmark for battery life prediction. In Proceedings of the Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, 2025, KDD '25, pp. 5789–5800. https://doi.org/10.1145/3711896.3737372.
26. Attia, P.M.; Grover, A.; Jin, N.; Severson, K.A.; Marber, T.M.; Liao, W.; Huber, M.H.; Ermon, S.; Braatz, R.D.; Chueh, W.C. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature* **2020**, *578*, 397–402.
27. Okita, T. Excluding the Target Domain Improves Extrapolation: Deconfounded Hierarchical Physics Constraints, 2026, [arXiv:cs.LG/2605.07485].

28. Benny, Y.; Galanti, T.; Benaim, S.; Wolf, L. Evaluation Metrics for Conditional Image Generation. *International Journal of Computer Vision* **2021**, *129*, 1712–1731. https://doi.org/10.1007/s11263-020-01424-w.
29. Mirza, M.; Osindero, S. Conditional generative adversarial nets. In Proceedings of the arXiv preprint arXiv:1411.1784, 2014.
30. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. In Proceedings of the Advances in Neural Information Processing Systems, 2015, Vol. 28.
31. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* **2022**.
32. Fawaz, H.I.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery* **2020**, *34*, 1936–1962.
33. Koochali, A.; Walch, M.; Thota, S.; Schichtel, P.; Dengel, A.; Ahmed, S. Quantifying quality of class-conditional generative models in time-series domain. *Applied Intelligence* **2023**, *53*, 23019–23038.
34. Kansal, R.; Li, J.; Parise, B.; Duarte, J.; Nachman, B. Evaluating generative models in high energy physics. *Physical Review D* **2023**, *107*, 076017.
35. Good, P.I. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd ed.; Springer, 2000.
36. Spearman, C. The Proof and Measurement of Association between Two Things. *American Journal of Psychology* **1904**, *15*, 72–101.
37. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall, 1994.
38. Mann, H.B.; Whitney, D.R. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics* **1947**, *18*, 50–60.
39. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *Journal of Machine Learning Research* **2012**, *13*, 723–773.

Appendix A Theoretical Analysis: Detection Power in the Small-Sample Regime

Section 3.3 states that the Fréchet distance requires distributional differences of order $\Omega(d/\sqrt{n_s})$ for detection, while MMD requires only $\Omega(1/\sqrt{n_s})$. This appendix provides the full derivations underlying those claims, including proofs of the two propositions, the ESS reliability threshold, and the complementarity regime that determines when each metric is preferable. These results formalize the design rationale for Layer 3 of the diagnostic framework: when per-stratum sample sizes are small relative to the feature dimension, S-MMD should supplement or replace SFD.

Appendix A.1 Detection Power Bounds

The Fréchet distance between two Gaussian distributions is computed from their means and covariance matrices. When these are estimated from finite samples, estimation error introduces noise into the distance computation. The critical question for evaluation practice is: how large must the true distributional difference be for the empirical Fréchet distance to reliably detect it?

Consider a stratum containing n_s real samples and n_s generated samples drawn from d-dimensional distributions P and Q. The variance of the empirical Fréchet distance $\hat{d}_F$ satisfies $\mathrm{Var}(\hat{d}_F) = \Omega(d^2/n_s)$, because estimating a $d \times d$ covariance matrix from n_s samples incurs Frobenius-norm error of order $O(d^2/n_s)$, and the matrix square root in the FD formula amplifies this error. For $\hat{d}_F$ to detect a distributional difference δ with probability exceeding $1/2$, the signal must exceed the noise, yielding the detection threshold $\delta^{FD}_{\min} = \Omega(d/\sqrt{n_s})$. By contrast, the biased MMD^2 estimator with a bounded kernel has variance $O(1/n_s)$ independently of d[39], giving $\delta^{MMD}_{\min} = \Omega(1/\sqrt{n_s})$.

Proposition A1 (Detection power ratio)**.** *The ratio of detection thresholds satisfies* $\delta_{\min}^{FD}/\delta_{\min}^{MMD} = \Omega(d)$.

Proof. Dividing the two thresholds: $\delta_{\min}^{FD}/\delta_{\min}^{MMD} = \Omega(d/\sqrt{n_s}) \cdot \sqrt{n_s} = \Omega(d)$. The n_s terms cancel, leaving a ratio that depends only on the feature dimension. □

The practical implication is immediate. For the 17-dimensional hand-crafted features used in our experiments, the Fréchet distance requires distributional differences roughly an order of magnitude larger than MMD to achieve comparable detection power. This does not mean that the Fréchet distance is inferior in all settings: when sample sizes are sufficient, FD captures covariance mismatch that MMD cannot. The point is that in the small-sample regime typical of battery degradation data, MMD provides a more reliable baseline.

Appendix A.2 Optimal Feature Dimension

Proposition 1 suggests that reducing the feature dimension d improves the detection power of the Fréchet distance. At the same time, reducing d discards information about the degradation curves, increasing the approximation error. This trade-off admits an optimal feature dimension that balances information loss against estimation stability.

Let $\epsilon_{\text{approx}}(d)$ denote the information loss from projecting to d dimensions (e.g., via PCA) and $\epsilon_{\text{est}}(d, n_s) = O(d/\sqrt{n_s})$ the estimation error from Proposition 1. The total error $\epsilon_{\text{total}}(d) = \epsilon_{\text{approx}}(d) + \epsilon_{\text{est}}(d, n_s)$ is minimized at d^*.

Proposition A2 (Optimal feature dimension)**.** *If* $\epsilon_{approx}(d) = O(d^{-\alpha})$ *for some* $\alpha > 0$ *(power-law decay of PCA residuals), then* $d^* = O(n_s^{1/(2\alpha+2)})$. *For* $\alpha \approx 1$ *(typical battery degradation features, where the eigenvalue spectrum decays roughly as* $1/k$*), this yields* $d^* = O(n_s^{1/4})$.

Proof. Setting $\partial\epsilon_{\text{total}}/\partial d = 0$ with $\epsilon_{\text{approx}} = c_a d^{-\alpha}$ and $\epsilon_{\text{est}} = c_e d/\sqrt{n_s}$ gives $d^* = (\alpha c_a \sqrt{n_s}/c_e)^{1/(\alpha+1)} = O(n_s^{1/(2\alpha+2)})$. □

This result provides a concrete guideline for practitioners. For the NASA 43°C stratum ($n_s = 4$), the optimal dimension is $d^* \approx 4^{1/4} \approx 1$–$2$. For the SNL 25°C stratum ($n_s = 34$), it is $d^* \approx 34^{1/4} \approx 2$–$3$. Both values are far below the full 17-dimensional feature space, indicating that aggressive dimension reduction is not merely advisable but necessary for reliable FD-based evaluation in these strata. Our S-MMD experiments use PCA with $d = 5$, which is a conservative choice that retains more information at the cost of some estimation stability.

Appendix A.3 ESS Threshold

The coefficient of variation (CV) of the FD estimator—the ratio of its standard deviation to its expected value—determines whether a given SFD estimate is practically useful. For the estimate to be meaningful, we require $\text{CV}(\hat{d}_F) < \gamma$ for some tolerance γ (e.g., $\gamma = 1$ means the standard deviation does not exceed the estimate itself).

From the variance bound in Section A.1, $\text{CV}(\hat{d}_F) \approx c \cdot d/\sqrt{n_s \cdot \delta}$, where δ is the true FD. Setting CV $< \gamma$ and solving for n_s yields $n_s > c^2 d^2/(\gamma^2 \delta^2)$. Expressing in terms of ESS:

$$\text{ESS} = \frac{n_s}{d+1} > \frac{c^2}{\gamma^2 \delta^2}$$

For moderate distributional differences ($\delta \sim O(1)$) and $\gamma = 1$, this gives ESS $\gtrsim 3$, consistent with our empirical guideline.

Two aspects of this result deserve emphasis. First, the required ESS depends on the true FD value δ: large distributional differences can be detected even at low ESS, while detecting small differences requires high ESS. This explains why Experiment 1 (where confusion produces substantially large FD values) yields clear results even for strata with ESS < 1, while Experiment 2 (where CVAE quality differences are subtle) is more sensitive to ESS. Second, the bootstrap confidence intervals reported in Section V-C confirm this theoretical prediction: CI width correlates with ESS across all strata and datasets.

Appendix A.4 Complementarity Regime

Combining the results above, we identify two regimes that determine which metric a practitioner should prioritize.

In the **large-sample regime** ($n_s \gg d^2$, equivalently ESS $\gg d$), both FD and MMD provide reliable estimates. FD is preferred in this regime because it captures both mean shift and covariance mismatch, providing richer diagnostic information than MMD.

In the **small-sample regime** ($n_s < d^2$, equivalently ESS $< d$), FD estimation becomes unreliable due to the covariance estimation noise that dominates the signal. MMD-based evaluation is preferred in this regime because its variance is independent of d.

The transition between regimes occurs at the ESS threshold $\tau(\delta)$ derived in Section A.3. The practical consequence is that a practitioner should compute ESS for each stratum and consult S-MMD whenever ESS falls below the threshold. When SFD and S-MMD agree on which strata are problematic (high ConsI), the diagnosis is robust regardless of the regime. When they disagree, the practitioner should give precedence to S-MMD for low-ESS strata and to SFD for high-ESS strata.

Appendix B SFD Algorithm

Algorithm A1 summarizes the computation of Stratified Fréchet Distance as described in Section III-A. The input consists of real and generated data, each annotated with a stratification variable s (e.g., temperature condition, temporal segment, or their cross-product). The algorithm partitions both datasets by stratum, computes the Fréchet distance within each stratum, averages the results (Within-SFD), and optionally adds a weighted pooled distance (Between-SFD). When $\lambda = 0$, the output reduces to the average per-stratum Fréchet distance, which is equivalent to CFID.

Algorithm A1 Stratified Fréchet Distance

Require: Real data $\{(x_i, s_i)\}$, generated data $\{(\hat{x}_j, s_j)\}$, stratification variable s, weight λ
1: Stratum set $\mathcal{S} \leftarrow \text{unique}(\{s_i\})$
2: within $\leftarrow 0$
3: **for** $s \in \mathcal{S}$ **do**
4: $\quad X_s, \hat{X}_s \leftarrow$ features of real and generated data in stratum s
5: $\quad$ within $\leftarrow$ within $+ d_F(X_s, \hat{X}_s)$
6: **end for**
7: within $\leftarrow$ within$/|\mathcal{S}|$
8: between $\leftarrow d_F(\text{all real}, \text{all generated})$
9: **return** within $+ \lambda \cdot$ between

The computational cost of SFD is dominated by the Fréchet distance computation within each stratum. Computing the Fréchet distance requires the sample mean ($O(n_s d)$), the sample covariance matrix ($O(n_s d^2)$), and the matrix square root of the product of two $d \times d$ covariance matrices ($O(d^3)$). Because SFD performs $|\mathcal{S}|$ such computations, the total cost is $O(nd^2 + |\mathcal{S}| \cdot d^3)$, where n is the total number of samples. For the battery degradation

datasets used in this study, $|\mathcal{S}| = 3$–6 and $d = 17$, so the additional cost relative to a single FID computation is negligible.

CRC and CDR (Layer 2) require computing $\binom{K}{2}$ pairwise inter-condition distances. Each pairwise distance costs $O(n_k d)$ for the mean-vector Euclidean metric or $O(n_k^2)$ for MMD. The permutation test adds a factor of N_{perm} (typically 1000–2000). For $K = 6$ and $N_{\text{perm}} = 1000$, the total Layer 2 cost is approximately 15,000 pairwise distance computations, which completes in under one minute on the hardware described in Section IV-D.

Bootstrap confidence intervals (Section V-C) multiply the cost of all three layers by the number of bootstrap iterations (500 in our experiments). The full evaluation pipeline—SFD, CRC, CDR, S-MMD, and bootstrap CIs for all metrics—requires approximately 30 minutes for the complete set of experiments reported in this paper.

Appendix C Supplementary Experiments

This appendix reports two sets of supplementary experiments. The first set (Experiments 3, 4, 5, 7b, 8b) complements the Layer 1 results in Section V by examining SFD sensitivity and design choices. The second set (Experiments 11–14) complements the Layer 2 result in Section V (Experiment 10) by providing detailed validation of MOP, CRC, and CDR across multiple datasets, random seeds, and controlled conditions.

Layer 1 Supplementary Experiments

Experiments 3, 4, and 5 examine the sensitivity of SFD to the weighting parameter λ, its generalization across datasets, and its formal relationship to CFID, respectively. Experiment 7b extends Experiment 7 by comparing knee-point-adaptive temporal segmentation with the equal-interval splitting adopted in Experiment 7 (Section 5), providing sensitivity analysis for that design choice. Experiment 8b extends Experiment 8 by analyzing a second form of inter-condition confounding arising from cycle-count differences across conditions. These experiments provide supporting evidence for the design choices made in the framework but are not essential to the main conclusions.

Experiment 3 (sensitivity to λ).

The parameter λ in the SFD formulation (Eq. 3) controls the balance between Within-SFD (per-stratum quality) and Between-SFD (overall distributional consistency). We analyzed the effect of λ on detection performance using the 43°C→4°C confusion in the NASA dataset (Table A18). At $\lambda = 0$ (Within-SFD only, equivalent to CFID), the confused-to-normal ratio reaches 507 times, the highest detection sensitivity observed. As λ increases, the contribution of Between-SFD—which is identical to FID and therefore subject to dilution—causes the detection ratio to decrease gradually (327 times at $\lambda = 2.0$). All values of λ substantially outperform FID. Setting $\lambda = 0$ maximizes per-condition detection sensitivity but forgoes evaluation of inter-condition distributional consistency. In the revised framework, the inter-condition assessment is handled by CRC and CDR (Layer 2) rather than by the Between-SFD term, so we adopt $\lambda = 0$ as the recommended default.

Table A18. Sensitivity analysis of λ (NASA, 43°C→4°C confusion).

λ	SFD (normal)	SFD (confused)	Ratio
0.0	0.006	3.21	507 times
0.5	0.008	3.31	441 times
1.0	0.009	3.41	392 times
2.0	0.011	3.61	327 times

Experiment 4 (cross-dataset validation).

To confirm that the advantage of SFD_c is not specific to the NASA dataset, we performed analogous confusion simulations on the three sub-datasets of BatteryLife (Table A19). In each case, the rarest condition was selected as the confusion target, and the detection advantage DA was computed. The results are consistent with Experiment 1: the smaller the minority proportion, the larger the DA. At CALB (minority proportion 7.4%), DA = 17.4; at SNL (14.8%), DA = 11.4—both exceeding a tenfold advantage of SFD_c over FID. At MICH (47.5%, nearly balanced), DA = 1.76, reflecting the fact that when minority and majority are roughly equal in size, dilution is mild and FID itself retains some detection capability. This pattern confirms that the effectiveness of SFD_c depends on a structural property of the data—the degree of condition imbalance—rather than on any artifact of a specific dataset.

Table A19. Cross-dataset validation (BatteryLife).

Dataset	N	Minority	Confusion	Proportion	DA
CALB	27	25°C	25°C→35°C	7.4%	17.4
SNL	61	15°C	15°C→25°C	14.8%	11.4
MICH	40	45°C	45°C→25°C	47.5%	1.76

Experiment 5 (relationship with CFID).

As noted in Section III, SFD_c with $\lambda = 0$ is mathematically equivalent to CFID, and we confirmed numerically that the difference is exactly zero. Table A20 shows the detection ratio (Model D / Model A) as λ varies. At $\lambda = 0$ (CFID-equivalent), the ratio is 1.99 times, substantially exceeding FID's 1.14 times. As λ increases and the Between-SFD term gains influence, the ratio decreases to 1.62 times at $\lambda = 1.0$. It remains above FID at every value of λ.

The per-condition FD breakdown provides informative detail. The excluded conditions 15°C and 35°C show deterioration of 2.72 times and 1.89 times respectively, while 25°C—the sole training condition—improves to 0.90 times. This per-condition profile makes it immediately apparent that Model D's quality varies drastically across conditions, information that is entirely lost in the single FID score.

Table A20. SFD_c vs. CFID (SNL, CVAE Model D/A ratio).

Metric	D/A ratio
FID	1.14 times
CFID (= SFD_c, $\lambda = 0$)	1.99 times
SFD_c ($\lambda = 0.25$)	1.85 times
SFD_c ($\lambda = 0.5$)	1.75 times
SFD_c ($\lambda = 0.75$)	1.68 times
SFD_c ($\lambda = 1.0$)	1.62 times

Experiment 7b (knee-point adaptive segmentation).

We compared four temporal segmentation strategies for $SFD_{c\times t}$: equal-interval $K = 2$ and $K = 3$, and knee-point-adaptive $K = 2$ and $K = 3$. The knee point was detected as the position of maximum curvature (minimum of the smoothed second derivative). Three generators were tested: normal (no confusion), confused (full condition replacement), and partial-second (only the latter half of the target condition is replaced).

On the NASA dataset, the mean knee-point positions were 0.42 (4°C, std = 0.26), 0.46 (24°C, std = 0.22), 0.45 (43°C, std = 0.05), and 0.30 (44°C, std = 0.00). The large inter-battery variability at 4°C and 24°C means that adaptive splitting assigns markedly different segment lengths to batteries within the same temperature condition.

Table A21 reports the NASA results.

Table A21. Knee-point vs. equal-interval segmentation (NASA, $SFD_{c\times t}$).

Strategy	Normal	Confused	Partial 2nd
Equal $K=2$	5.39	828.3	819.5
Knee $K=2$	0.90	13.3	210.7
Knee $K=3$	1.05	9.0	182.6
Equal $K=3$	12.8	9.3	4.1

Equal $K = 2$ achieves the highest confused-to-normal ratio (153.7 times), while knee $K = 2$ achieves a lower baseline under normal generation (0.90 vs. 5.39). On the SNL dataset, all strategies produced similar SFD values, reflecting the fact that the confusion pattern (35°C → 25°C) affects the entire curve uniformly rather than a specific temporal phase.

Experiment 8b (cycle-count confounding).

A second form of inter-condition confounding arises from differences in cycle lifetime. When comparing degradation curves across batteries with different lifetimes, temporal normalization is required. The most common approach is fractional normalization, which expresses each point as a percentage of total lifetime. This normalization can conceal physically meaningful differences in time scale.

In the NASA dataset, the mean cycle lifetime at 24°C is 122 cycles, compared to only 40 cycles at 43°C. Under fractional normalization, the "50% point" corresponds to cycle 61 for 24°C but cycle 20 for 43°C. Data from physically distinct degradation stages are mapped onto the same normalized coordinate, obscuring the underlying difference.

To quantify this effect, we computed the FD ratio for a 43°C→24°C confusion under both fractional and physical (absolute cycle count) normalization. The FD ratio under fractional normalization is 3.66 times. Under physical normalization, it reaches 463,920 times (time-scale confounding ratio TCS = 126,724). This orders-of-magnitude discrepancy reveals that fractional normalization suppresses the time-scale confound arising from the 3.0-fold difference in cycle lifetime.

By contrast, in the SNL dataset, where cycle counts are approximately uniform across batteries (∼300 cycles), TCS ≈ 1.0 and the normalization strategy makes no difference. Two lessons emerge. First, when cycle lifetimes vary substantially across conditions, the choice of normalization can critically affect evaluation outcomes. Second, the SFD framework can serve as a tool for validating normalization choices post hoc, by computing SFD under different normalizations and examining whether the results diverge.

Layer 2 Supplementary Experiments

The following four experiments provide detailed validation of the MOP, CRC, and CDR indicators introduced in Section III and evaluated in Experiment 10.

Experiment 11 (MOP seed dependence, $n_{\mathbf{seeds}} = 10$). Experiment 10 reports the mean detection rate across 10 seeds. Experiment 11 examines whether this result is stable across individual seeds or driven by outliers. The concern is that a single favorable random initialization might produce a high MOP even for a defective generator, inflating the apparent detection rate. To test this, the same 10 CVAE models from Experiment 10 are analyzed seed by seed.

Table A22 shows that the normal reference MOP is tightly concentrated (std = 0.029), confirming that the projection axis $\mathbf{v}_1$ is consistently recovered across random initializations. Every individual seed detects type-2 inversion (MOP < 0) and type-3 confusion (MOP < 0.471). The type-4 uniform generator shows high variance (std = 0.330) because a condition-free generator occasionally reproduces partial temperature ordering by chance.

The Mann-Whitney test remains significant ($p < 0.0001$) because the *distribution* of MOP values is systematically lower than the normal reference even when individual seeds vary.

Table A22. Experiment 11: MOP across 10 random seeds (SNL 1C+2C, $K = 6$, $d = 5$). Mann-Whitney U test compares each defective generator's MOP distribution against the normal reference.

Generator	MOP mean	MOP std	p-value	Detection
Normal (reference)	$+0.971$	0.029	—	—
Type-2 inversion	-0.400	0.118	< 0.0001	100%
Type-3 confusion	$+0.406$	0.028	< 0.0001	100%
Type-4 uniform	$+0.109$	0.330	< 0.0001	100%

Experiment 12 (MOP cross-dataset generalization, $n_{\text{seeds}} = 10$). Experiment 10 evaluates MOP on the SNL dataset. Experiment 12 asks whether the same approach generalizes to the other three battery datasets (NASA, CALB, MICH), each of which differs in sample size, number of conditions, and temperature range. The key dataset-level factor is projection quality ρ_{proj}. When the real data do not exhibit a clear temperature gradient in the feature space, the projection axis $\mathbf{v}_1$ cannot discriminate conditions reliably.

Table A23 shows that MOP generalizes fully to MICH ($\rho_{\text{proj}} = 1.000$) and NASA ($\rho_{\text{proj}} = 0.800$) for type-2 and type-3 detection. The CALB type-2 failure (0% detection, $\rho_{\text{proj}} = 0.200$) is not a deficiency of MOP but a dataset-level limitation. Cell-to-cell variability in CALB overwhelms the temperature signal in the feature space, making the projection axis unreliable. The NASA type-4 failure (CDR inflation) is explained by Experiment 14.

Table A23. Experiment 12: MOP cross-dataset generalization ($n_{\text{seeds}} = 10$). ρ_{proj}: Spearman correlation between projected real centroids and temperature labels.

Dataset	ρ_{proj}	Ref. MOP	Type-2	Type-3	Type-4
SNL	$+0.717$	$+0.971$	100%	100%	100%
NASA	$+0.800$	$+0.520$	100%	90%	0%†
CALB	$+0.200$	$+1.000$	0%‡	100%	100%
MICH	$+1.000$	$+1.000$	100%	100%	100%

†CDR inflation due to $n = 3$–7 per condition; see Experiment 14. $^{\ddagger}\rho_{\text{proj}} = 0.200 < 0.7$; projection axis does not capture temperature ordering.

Experiment 13 (CRC minimum K, random subsampling, $n = 100$ trials). The theoretical lower bound on the CRC permutation test p-value is $p_{\min} = 1/M!$ where $M = \binom{K}{2}$, giving $p_{\min} = 0.167$ at $K = 3$. This predicts that CRC cannot reach significance at conventional $\alpha = 0.05$ for small K regardless of how severe the defect is. Experiment 13 tests this prediction empirically by randomly subsampling K conditions from the full pool of 24 SNL dataset conditions and measuring the type-3 detection rate over 100 trials. Random subsampling is essential here because earlier analyses in Section 5 drew conditions sequentially, confounding K with the specific conditions selected.

Table A24 shows that detection exceeds 80% only at $K \geq 8$, consistent with the theoretical bound. The non-monotonicity at $K = 9$ (74%) reflects sampling variance across 100 trials rather than a genuine decrease in power. The practical recommendation is $K \geq 8$ for reliable CRC detection. When fewer conditions are available, MOP should be used instead.

Experiment 14 (root cause of CDR inflation). In Experiment 12, the NASA type-4 detection rate is 0%. The reason is that the normal CVAE generator itself yields CDR $= 1.985 \gg 1$, which would falsely suggest that the generator exaggerates inter-condition distances. Experiment 14 identifies the cause by varying one factor at a time in a controlled

Table A24. Experiment 13: CRC type-3 detection rate vs. K (random subsampling from SNL pool of 24 conditions, $n = 100$ trials per K).

K	4	5	6	7	8	9	10	12	15	18	24
Det.	25%	51%	59%	72%	80%	74%	81%	83%	90%	85%	94%

CVAE experiment on SNL data. Four settings are compared. Setting A uses standard training. Setting B overfits by training for 5× the normal epochs. Setting C scales condition labels by a factor of 10. Setting D restricts to $n = 2$ cells per condition to simulate data shortage.

Table A25 shows that only setting D (data shortage) inflates CDR. Overfitting (B) and label scaling (C) have negligible effect. This confirms that CDR inflation in NASA arises from the small per-condition sample size ($n = 3$–7), not from the model architecture or training procedure. The practical implication is that CDR should not be applied when $n < 10$ per condition. In such cases, type-4 detection must rely on visual inspection of generated curves.

Table A25. Experiment 14: CDR under four controlled conditions (SNL, $n_{\text{seeds}} = 5$). Each setting varies one factor from the standard training procedure.

Setting	Description	CDR mean	CDR std
A (normal)	Standard training	0.947	0.035
B (overfit)	Epochs ×5	0.981	0.010
C (scale)	Condition label ×10	0.996	0.028
D (shortage)	$n = 2$ per condition	3.227	0.063

Appendix D Diagnostic Guide: Applicability Conditions and Worked Example

This appendix synthesizes the empirical findings from Sections 5.2 and Appendix C into a practical reference for applying the three-layer framework. The first part establishes the defect taxonomy and the mapping from defect type to detection indicator. The second part describes the conditions under which each indicator produces reliable results and the conditions under which it does not, together with supporting evidence from the validation experiments. The third part gives the decision procedure. The fourth part illustrates the procedure on the HPC-FNO-CFM evaluation reported in Experiment 9.

Appendix D.1 Defect Taxonomy and Indicator Mapping

The framework targets four types of structural defect in conditional generative models, each of which is invisible or poorly localized by FID.

A **type-1 (degenerate)** generator collapses to a single mode, reproducing the same output regardless of the conditioning variable. Effective Stratum Size (ESS) below one signals this failure.

A **type-2 (inverted)** generator reproduces physically plausible curves at each condition, but with the temperature ordering reversed. MOP below zero signals this failure.

A **type-3 (confused)** generator mixes adjacent conditions, failing to distinguish, for example, the degradation behavior at 25°C from that at 35°C. MOP below the dynamic threshold $\mu_{\text{normal}} - 0.5$ signals this failure. CRC provides supplementary confirmation when $K \geq 8$ conditions are available.

A **type-4 (uniform)** generator loses condition specificity entirely, producing all conditions from effectively the same distribution. CDR substantially below unity signals this failure.

Table A26 summarizes the indicator mapping.

Table A26. Defect types and primary detection indicators. The auxiliary column lists indicators that provide corroborating evidence when the primary indicator is inconclusive.

Defect type	Primary indicator	Auxiliary
Type-1 degenerate	ESS < 1	—
Type-2 inverted	MOP < 0	—
Type-3 confused	MOP $< \mu_{\text{normal}} - 0.5$	CRC
Type-4 uniform	CDR $\ll 1$	—

Appendix D.2 Applicability Conditions

Each indicator requires certain properties of the evaluation data. When these properties do not hold, the indicator may produce unreliable results regardless of the quality of the generative model. Table A27 provides a consolidated reference. The subsections below discuss each indicator in turn.

For CRC, the indicator requires at least $K = 3$ conditions for the permutation test [35] to be computable. However, the theoretical minimum achievable p-value at $K = 3$ is $1/M! = 1/6 = 0.167$ where $M = \binom{K}{2}$, which exceeds the conventional significance level of 0.05. Experiment 13 confirmed empirically by random subsampling from the SNL dataset (100 trials per K) that detection rates exceed 80% only when $K \geq 8$. In addition, real battery degradation data often exhibit non-monotone inter-condition distance matrices due to cell-to-cell variability, which prevents CRC from distinguishing genuine type-2 inversion from natural data variation. CRC is therefore best used as supplementary confirmation of type-3 confusion rather than as the primary detection mechanism for type-2 inversion.

For MOP, the indicator requires the temperature axis to be recoverable from the real data feature space. Projection quality is measured as

$$\rho_{\text{proj}} = \rho_S\Big(\{\pi_k^{\text{real}}\}_{k=1}^{K},\ \{t_k\}_{k=1}^{K}\Big),$$

the Spearman rank correlation [36] between projected per-condition centroids and normalized temperature labels. Table A28 reports MOP detection rates across the four datasets used in this study, together with the projection quality for each.

On the SNL, NASA, and MICH battery datasets, ρ_{proj} exceeds 0.7, and MOP achieves 100% type-2 detection on three of these. On the CALB battery dataset, $\rho_{\text{proj}} = 0.200$: cell-to-cell variability overwhelms the temperature signal in the feature space, and type-2 detection is not achieved. This reflects a property of the dataset, not a limitation of MOP.

For CDR, the indicator requires $n \geq 10$ samples per condition. At smaller sample sizes, the mean-vector distance estimate is dominated by sampling noise. Experiment 14 confirmed this by showing that reducing the per-condition sample count to $n = 2$ inflates CDR to 3.2 for a correctly conditioned generator, while overfitting the generator or scaling the condition labels has negligible effect. Table A29 reports CDR type-4 detection rates across the four datasets.

On the NASA battery dataset, the reference generator yields $\text{CDR}_{\text{ref}} = 1.985 \gg 1$ due to the small per-condition sample count, confirming that CDR is not applicable for that dataset.

A key property of the Layer-2 indicators is that MOP and CDR are mutually non-sensitive. Type-2 and type-3 defects do not trigger CDR, because inter-condition distances

Table A27. Required conditions and identified failure modes for each indicator. Failure modes were identified through controlled experiments (Experiments 11–14, Appendix C).

Indicator	Required conditions	Failure mode
ESS	No special requirements	Below one: covariance estimate unreliable; switch to S-MMD
CRC	$K \geq 8$ (randomly sampled conditions); real data distance matrix is monotone	$K < 8$: minimum achievable p-value exceeds 0.05; non-monotone matrix: type-2 inversion detection unreliable
MOP	$\rho_{\text{proj}} \geq 0.7$; $n_{\text{seeds}} \geq 5$	$\rho_{\text{proj}} < 0.5$: type-2 detection degrades (0% on CALB at $\rho_{\text{proj}} = 0.200$); insufficient seeds: MOP distribution estimate unreliable
CDR	$n \geq 10$ per condition	$n < 10$: mean-vector distance dominated by sampling noise; CDR inflated to 3.2 at $n = 2$ (Experiment 14)

Table A28. MOP detection rates across the four battery datasets ($n_{\text{seeds}} = 10$, Experiments 10–12). ρ_{proj}: projection quality. Ref. MOP: mean MOP of the correctly conditioned reference generator. Type-2 and type-3 columns report the detection rate for the corresponding defect. Type-4 detection is handled by CDR, not MOP.

Dataset	ρ_{proj}	Ref. MOP	Type-2	Type-3
SNL battery dataset	0.717	+0.971	100%	100%
NASA battery dataset	0.800	+0.140	100%	90%
CALB battery dataset	0.200	+1.000	0%	100%
MICH battery dataset	1.000	+1.000	100%	100%

are preserved even when their ordering is wrong. Type-4 defects do not consistently trigger MOP, because a condition-free generator may accidentally reproduce partial ordering by chance. This complementarity is confirmed experimentally on the SNL battery dataset in Table A30, and means that a positive result from either indicator is sufficient to flag the corresponding defect class, and a negative result from one does not rule out defects of the type that the other indicator targets.

Table A31 consolidates the effective indicators and detection rates for each of the four battery datasets. Data-rich datasets (MICH, SNL) support all Layer-2 indicators. Data-scarce datasets (NASA)[1] require CDR to be omitted and MOP results to be interpreted with awareness that the reference MOP is lower than on other datasets (+0.140 vs. +0.971 on SNL).

Appendix D.3 Decision Procedure

The following procedure applies the three-layer framework sequentially. Each step either identifies a defect and concludes the diagnostic, or confirms structural integrity and proceeds to the next step.

Pre-evaluation checks

Before computing any indicator, verify two dataset-level properties from the real data alone. First, the per-condition sample count n: if $n < 10$ for any condition, CDR is not applicable for that condition. Second, the projection quality ρ_{proj}: if $\rho_{\text{proj}} < 0.5$, MOP sensitivity is degraded and results should be interpreted with caution. Both quantities can be computed before any generative model is evaluated.

[1] The terms *data-rich* and *data-scarce* here refer specifically to the number of cells available *per operating condition*, not to the total number of data points in the dataset. NASA is one of the most widely used battery degradation benchmarks and contains a substantial number of charge–discharge cycles per cell. However, it covers only a small number of distinct temperature conditions, and the number of cells tested at each condition is small (3–7 cells per condition). It is this *per-condition cell count* that determines whether CDR, CRC, and MOP can be applied reliably—not the total dataset size. By contrast, the MICH battery dataset has approximately 20 cells per condition, placing it comfortably above the $n \geq 10$ threshold required for CDR.

Table A29. CDR type-4 detection rates across the four battery datasets (Experiments 1, 2, 9). CDR_{ref}: CDR of the correctly conditioned reference generator. CDR is omitted for datasets where $n < 10$ per condition.

Dataset	n per cond.	CDR_{ref}	Type-4	CDR valid
SNL battery dataset	4–12	0.947	100%	Yes
NASA battery dataset	3–7	1.985	0%	No
CALB battery dataset	∼7	0.972	100%	Yes
MICH battery dataset	∼20	0.997	100%	Yes

Table A30. Indicator cross-sensitivity on the SNL battery dataset ($n_{seeds} = 5$). A checkmark indicates that the indicator detects the defect type. CRC requires $K \geq 8$ conditions ($*$). CDR may spuriously increase for type-4 defects when sample sizes are very small ($**$).

Defect type	MOP	CDR	CRC
Type-2 inverted	◯	×	×
Type-3 confused	◯	×	◯*
Type-4 uniform	×	◯	×**

Layer 1

Compute $\mathrm{ESS}(s_k) = n_k/(d+1)$ for each condition stratum. Strata with ESS < 1 have rank-deficient covariance matrices; use S-MMD as a covariance-free alternative for those strata and treat Fréchet distance estimates as order-of-magnitude indicators rather than precise measurements.

Layer 2

Train $n_{seeds} \geq 5$ independent instances of the generative model. Compute MOP for each instance and apply a one-sided Mann-Whitney U test [38] comparing the resulting distribution against that of a correctly conditioned reference generator. If the mean MOP falls below zero with $p < 0.05$, report type-2 inversion. If the mean MOP falls below $\mu_{\text{normal}} - 0.5$ with $p < 0.05$, report type-3 confusion. When $K \geq 8$ and the real distance matrix is monotone, compute CRC to corroborate a type-3 finding. When $n \geq 10$ per condition, compute CDR and construct a bootstrap confidence interval [37]. If CDR falls substantially below unity with a confidence interval that excludes unity, report type-4 uniform collapse.

Layer 3

Compute ConsI, the Spearman rank correlation [36] between per-stratum SFD and per-stratum S-MMD. High ConsI corroborates the Layer-1 diagnosis. Low ConsI indicates that Fréchet distance estimates may be distorted by small sample sizes in some strata, and the practitioner should weight the Layer-2 indicators more heavily.

Figure A5 illustrates the procedure as a flowchart.

Table A31. Effective indicators and detection rates by dataset ($n_{\text{seeds}} = 10$, Experiments 10–12). T2, T3, T4: type-2, type-3, type-4 detection rates. CDR valid: whether CDR meets the $n \geq 10$ per condition applicability condition.

Dataset	n/cond.	MOP T2	MOP T3	CDR T4	CDR valid	Notes
SNL battery dataset	4–12	100%	100%	100%	Yes	All indicators applicable
NASA battery dataset	3–7	100%	90%	0%	No	CDR not applicable; CDR_{ref} inflated to 1.985
CALB battery dataset	∼7	0%	100%	100%	Yes	Type-2 not detected; $\rho_{\text{proj}} = 0.200 < 0.7$
MICH battery dataset	∼20	100%	100%	100%	Yes	All indicators applicable

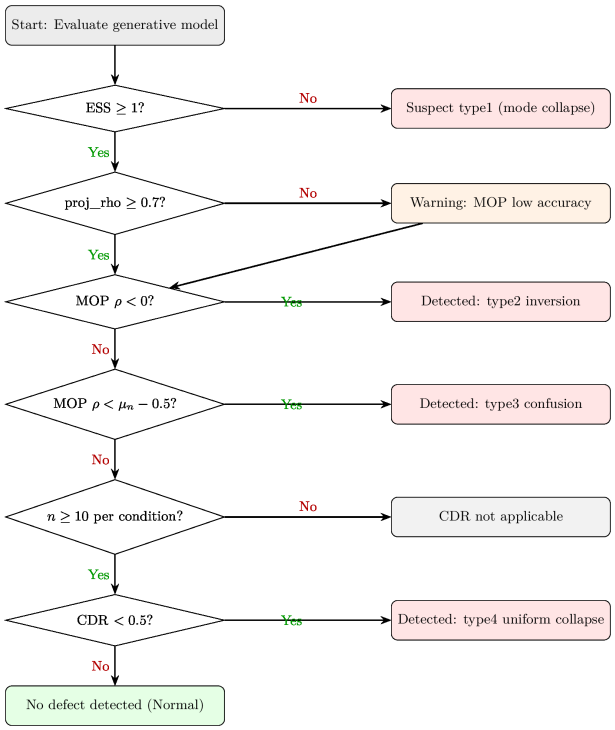

Figure A5. Decision procedure for the three-layer diagnostic framework. ρ_{proj} denotes the projection quality of the MOP axis. μ_{normal} denotes the mean MOP of the reference generator. Conditions labeled "not applicable" indicate that the indicator cannot produce a reliable result for the given dataset, reflecting a property of the evaluation data rather than of the generative model.

Appendix D.4 Worked Example: HPC-FNO-CFM Evaluation

This section illustrates the decision procedure by applying it to the HPC-FNO-CFM evaluation reported in Experiment 9 (Section V-A). The model is a conditional Flow Matching model integrating a Fourier Neural Operator, evaluated on the NASA battery dataset across five temperature conditions (4, 8, 13, 24, 43°C). Five variants are compared, differing in how many layers of the pre-trained FNO are retained during fine-tuning.

The worked example serves two purposes. It shows what each indicator detected in this evaluation. It also shows what each indicator could not detect, and what data conditions would be required for it to detect the corresponding defect. The second purpose is at least as important as the first for a practitioner deciding whether to apply this framework to a new dataset.

Pre-evaluation checks

The NASA battery dataset provides $n = 3$–15 cells per temperature condition. For conditions with $n < 10$, CDR cannot produce reliable inter-condition distance estimates

Table A32. Three-layer diagnostic results for the five HPC-FNO-CFM variants on the NASA battery dataset (Experiments 9 and 10). SFD is the mean per-condition Fréchet distance (lower is better). CRC is the Spearman rank correlation of inter-condition distances (higher is better; range $[-1, +1]$). CDR is not reported because $n < 10$ per condition on the NASA battery dataset. MOP ρ is the Spearman rank correlation between projected per-condition centroids of the real and generated data ($\rho > 0$: ordering preserved; $\rho < 0$: ordering inverted). MOP diagnosis applies the thresholds established in Experiment 10.

Model variant	SFD	CRC	MOP ρ	MOP diagnosis
`freeze_2layer`	2,353	$+0.724$	-0.800	**Type-2 inversion** ($\rho < 0$)
`freeze_3layer`	41,747	-0.005	$+1.000$	No defect; CRC anomaly
`freeze_1layer`	4.9×10^6	$+0.418$	$+0.500$	No defect detected
`scratch_3layer`	1.0×10^9	$+0.123$	$+0.300$	No defect detected
`pure_cfm`	1.8×10^{10}	$+0.033$	$+0.700$	No defect detected

because the per-condition mean-vector distance is dominated by sampling noise at this sample size. In the controlled experiments of Experiment 14, a correctly conditioned CVAE generator yields CDR $= 3.2$ at $n = 2$, which would falsely indicate distance overestimation.[2] CDR is therefore not applicable to the NASA battery dataset at its current sample size. To make CDR applicable, at least $n = 10$ cells per condition are required. The current evaluation has 3–7 cells for most conditions.

The projection quality is $\rho_{\text{proj}} = 0.900$, which exceeds the applicability threshold of 0.7. MOP is therefore applicable.

Layer 1

With feature dimension $d = 50$ and per-condition sample counts of 3–15, the Effective Stratum Size is ESS $= n/(d+1) \leq 15/51 \approx 0.29$ for all conditions. All strata have ESS < 1, indicating that the covariance matrices used in Fréchet distance computation are rank-deficient. SFD values for this dataset should therefore be treated as order-of-magnitude indicators rather than precise measurements. This is confirmed by the S-MMD comparison in Experiment 9. Despite a seven-order-of-magnitude range in SFD across the five variants, S-MMD values vary only between 0.22 and 0.31, indicating that the Fréchet distance differences reflect covariance estimation noise at least as much as genuine quality differences.

The S-MMD values do distinguish the variants, though with much smaller separation than SFD. In the low-ESS regime, Layer-2 indicators that operate on mean vectors rather than covariance matrices provide more stable evidence.

Layer 2

Table A32 reports the full three-layer diagnostic results for each variant, including the MOP value and diagnosis.

The SFD values span seven orders of magnitude. `freeze_2layer` achieves the lowest mean per-condition Fréchet distance (2,353), while `pure_cfm` reaches 1.8×10^{10} with catastrophic failure at 8°C. This ordering reflects the degree of physical constraint imposed by the pre-trained FNO layers.

[2] As defined in the footnote accompanying Experiment 14, CDR $\gg 1$ is termed *distance overestimation*: the generated data place conditions further apart than the real data do. At $n = 2$ per condition, the per-condition mean-vector estimates are dominated by individual cell variability, causing the estimated inter-condition distance to fluctuate far from its true value. The resulting CDR inflation is a consequence of the small sample size, not of any property of the generative model.

CRC reveals the structural integrity of each variant. `freeze_2layer` achieves the highest CRC (+0.724), indicating that the ordering of inter-condition distances is substantially preserved. `pure_cfm` achieves CRC = +0.033, indicating that the physical structure across temperature conditions has been almost entirely destroyed, even if individual curves may appear plausible. `freeze_3layer` achieves CRC = −0.005: the inter-condition distance ordering is essentially random, despite this variant achieving perfect temperature ordering (MOP = +1.000). This combination—correct ordering but scrambled distances—is itself a diagnostically meaningful finding. The variant preserves the direction of the temperature effect but not its magnitude structure.

MOP detects a type-2 inversion defect in `freeze_2layer` (MOP = −0.800 < 0). This means that `freeze_2layer` reproduces degradation curves in the wrong temperature order. The configuration with two frozen FNO layers appears to reverse the direction of the temperature extrapolation. `freeze_3layer` achieves MOP = +1.000, confirming that freezing three layers preserves the temperature ordering correctly, even though the absolute curve quality is lower (SFD = 41,747).

The CRC result requires $K = 5$ conditions for the NASA battery dataset, yielding $M = \binom{5}{2} = 10$ condition pairs. The theoretical minimum p-value at $K = 5$ is $1/10! < 10^{-6}$, so CRC can in principle achieve significance at this condition count. However, the empirical detection rate from Experiment 13 shows that $K \geq 8$ is needed for reliable detection (80% rate) under random subsampling. With $K = 5$, the CRC results in Table A32 should be interpreted as directional evidence rather than statistically confirmed findings. To achieve reliable CRC detection on this dataset, additional temperature conditions would need to be included in the experimental design, bringing K to at least 8.

What the evaluation could not determine

CDR was not applicable because $n = 3$–7 cells per condition falls below the required $n \geq 10$. As a result, type-4 uniform collapse could not be formally assessed for any of the five variants. The CDR values observed in Experiment 9, which ranged from 4 to 958 across condition pairs, are inflated by sampling noise rather than reflecting genuine distance overestimation. To make CDR applicable to the NASA battery dataset, approximately 10 cells per temperature condition would be required, compared to the 3–7 currently available.

For the T/J confounding detection, the NASA battery dataset has a C-rate effect size of approximately 1.25, for which Experiment 8 indicates that approximately $n = 8$ cells per condition are needed for the T/J ratio to achieve 80% detection. The current per-condition counts are near this threshold, so confounding detection is marginally feasible on the NASA battery dataset, unlike on the SNL battery dataset where the C-rate effect size of 0.426 would require approximately 50 cells per condition—roughly four times the available count.

Layer 3

The near-uniform S-MMD values (0.22–0.31) across the five variants confirm low consistency between SFD and S-MMD for this dataset. The Layer-2 indicators (CRC, MOP) are therefore the primary diagnostic evidence for this evaluation.

Summary

Table A33 summarizes what each indicator detected, what it could not detect, and what data conditions would enable it.

The most important diagnostic finding from this evaluation is the type-2 inversion in `freeze_2layer`. This model achieves the best per-condition generation quality (lowest SFD) and the best inter-condition distance ordering (highest CRC), yet it reproduces the temperature conditions in the wrong order (MOP = −0.800). This combination—high

Table A33. Summary of diagnostic findings and data requirements for the HPC-FNO-CFM evaluation on the NASA battery dataset. The "requirement to enable" column states the data condition that would allow the indicator to produce reliable results on this dataset.

Indicator	What was detected	What could not be determined	Requirement to enable
SFD	Seven-order-of-magnitude quality range. `freeze_2layer` achieved the lowest SFD, `pure_cfm` the highest	Precise per-condition estimates (ESS $<$ 1 for all strata)	$n \geq 51$ per condition (ESS ≥ 1 at $d = 50$)
CRC	`freeze_2layer` best structural ordering (CRC $= 0.724$); `pure_cfm` almost entirely destroyed (CRC $= 0.033$)	Statistically confirmed type-3 detection ($K = 5 < 8$)	$K \geq 8$ temperature conditions in the experimental design
CDR	Not applicable	Type-4 uniform collapse for any variant	$n \geq 10$ cells per temperature condition
MOP	Type-2 inversion in `freeze_2layer` (MOP $= -0.800$); correct ordering in `freeze_3layer` (MOP $= +1.000$)	Statistical significance requires $n_{\text{seeds}} \geq 5$ independent model runs	Already met ($\rho_{\text{proj}} = 0.900 \geq 0.7$)

quality with inverted ordering—is a failure mode that neither FID nor CRC alone would reveal. FID cannot localize the inversion, and CRC measures distance ordering rather than temperature ordering. MOP is the only indicator in the framework that directly addresses this property, and this evaluation provides a concrete example of why it is necessary.

The evaluation also illustrates the data requirements for comprehensive diagnostics. With the current NASA battery dataset (5 conditions, 3–7 cells per condition), only SFD, CRC, and MOP can be applied. CDR and statistically confirmed CRC-based type-3 detection would require a redesigned experimental protocol with more cells per condition and more temperature conditions. These are conclusions about experimental design, not about the quality of the generative models under evaluation.