

Unsupervised Conformal Novelty Detection for Hierarchical Data

Sieun Kim¹ and Seonghun Cho^{*1}

¹Department of Statistics, Inha University, Incheon, South Korea

Abstract

Electric vehicle (EV) battery packs exhibit a natural pack–module–cell hierarchy, which induces dependence among measurements within the same module. Such hierarchical dependence poses challenges for the direct application of conventional novelty detection methods. To address these challenges, we develop conformal e-value procedures for hierarchical novelty detection, with the goal of controlling the false discovery rate (FDR) at both the group and unit levels. We combine hierarchical conformal score construction with eBH and U-eBH multiple testing procedures, and consider split conformal, full conformal, and group-wise conformal regimes. The proposed methods are evaluated through simulation studies and an analysis of EV battery pack data, where they provide empirical FDR control and detect localized module- and cell-level irregularities.

Keywords: novelty detection, hierarchical structure, conformal inference.

1 Introduction

Anomaly detection, the task of identifying observations that deviate significantly from the established norm, is a fundamental problem across a wide range of scientific and industrial fields. Its applications are diverse, encompassing fraud detection in insurance, credit, and finance (Bolton & Hand, 2002), medical diagnostics through functional magnetic resonance imaging (fMRI) and brain wave recordings (Fernando et al., 2021), and complex system log analysis (He et al., 2016; Guo et al., 2021). In the industrial sector, anomaly detection is particularly crucial for machine health monitoring and defect identification (Kim et al., 2012; Bertolini et al., 2021; Bae et al., 2023; Liu et al.,

^{*}To whom all correspondence should be addressed. Email: seonghun.cho@inha.ac.kr

2024), where it serves as the backbone for automatic quality control and predictive maintenance.

A prominent example of high-stakes industrial anomaly detection is the quality assurance of electric vehicle (EV) battery packs. In this study, we analyze a comprehensive EV battery pack dataset (Ministry of SMEs and Startups & KAIST, 2022) provided by the Korea AI Manufacturing Platform (KAMP). KAMP is a specialized online platform organized by the Ministry of SMEs and Startups of South Korea to advance AI-driven manufacturing solutions. The dataset captures voltage and temperature measurements during the charge–discharge testing phase—a critical step in the assembly process to determine if a battery pack is acceptable or defective before being supplied for electric buses and commercial vehicles. The EV battery data presents a unique challenge due to its inherent hierarchical structure. Each battery pack is composed of 16 modules, each containing 11 cells. While voltage is measured at the individual cell level, temperature is recorded at the module level. This nested organization introduces complex dependencies among measurements within the same group, violating the standard independent and identically distributed (IID) assumption. Conventional novelty detection methods that ignore this grouping often suffer from a loss of statistical power or fail to provide rigorous error control, as they do not account for within-group correlations.

In this paper, we mainly consider group- and unit-level novelty detection. To address the aforementioned challenges, we propose a framework for hierarchical conformal novelty detection. Our approach integrates conformal inference with the e-value-based Benjamini-Hochberg (eBH) procedure (Wang & Ramdas, 2022) to achieve valid false discovery rate (FDR) control without making strong parametric assumptions. The primary contributions of this paper are two-fold. First, we formulate hierarchical conformal novelty detection procedures for both group-level and unit-level hypotheses, and implement them under three conformal regimes: hierarchical split conformal (HSC), hierarchical full conformal (HFC), and hierarchical group-wise conformal (HGC) novelty detection. These regimes reflect different trade-offs between data usage and outlier contamination. HSC avoids contamination by using only clean reference data for training, but it may lose power due to data splitting. HFC uses all available data, but it can be more sensitive to contamination from outliers in the test set. HGC repeats the learning process for each test group, isolating potential contamination while using more reference samples than HSC. Second, we provide the theoretical foundation for these hierarchical procedures. For the theoretically justified regimes, we prove that the proposed statistics are compound e-values, satisfying the average expectation condition required for valid FDR control via the eBH procedure. We also show that the

localized training in HGC preserves the hierarchical exchangeability needed for the validity of the group-level detection procedure.

The remainder of this paper is organized as follows. Section 2 establishes the foundational framework for conformal (group- and unit-level) novelty detection in hierarchical data. Specifically, we introduce the development of our three hierarchical conformal variants, alongside their respective theoretical guarantees. Section 3 provides a comparative analysis through numerical simulations to evaluate the statistical power and error control of the proposed methods. In Section 4, we demonstrate the practical utility of our framework by applying it to a real-world EV battery pack dataset. Finally, Section 5 offers concluding remarks and discusses potential directions for future research. For completeness, all technical proofs are deferred to the appendix.

2 Conformal novelty detection for hierarchical data

In novelty detection problems, analysts may focus on either group-level or unit-level detection, depending on the specific context of their study. We describe the formal setup for these two problems, propose new conformal detection methods, and establish their validity under minimal assumptions.

In the remainder of the paper, we adopt the following notation. For an integer $n \in \mathbb{N}$, $[n]$ denotes the index set $\{1, \dots, n\}$, and $\mathcal{S}_n = \{\sigma : [n] \rightarrow [n], \sigma \text{ is a bijection}\}$ represents the set of all permutations on $[n]$. For $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$. For a finite set M , $|M|$ denotes its cardinality.

2.1 Problem setup

We consider the task of hierarchical novelty detection. We are provided with two datasets: a reference set consisting of samples from n groups, denoted by $\mathcal{D}_{\text{ref}} = \{X_{i,j} : i \in [n], j \in [K]\}$, and a test set from m groups, denoted by $\mathcal{D}_{\text{test}} = \{X_{n+i,j} : i \in [m], j \in [K]\}$. Here, each group contains a fixed number of units. We let $X_{i,j} \in \mathcal{X}$ ($i \in [n+m], j \in [K]$) denote the measurement of the j -th unit in the i -th group, and let $Z_i = (X_{i,1}, \dots, X_{i,K}) \in \mathcal{X}^K$ be the vector of K measurements within the i -th group. While the assumption of a fixed group size K is more restrictive than the random group size settings in Lee et al. (2023) and Lee & Ren (2025), we adopt it here as it is frequently encountered in industrial manufacturing applications.

In the novelty detection framework, the reference set is assumed to contain only *inliers*, whereas the test set may contain a mixture of *inliers* and *outliers*. The hierar-

chical dependence structure among the samples is formally defined as follows:

$$\begin{aligned}
P_{0,1}, \dots, P_{0,n+m} &\stackrel{\text{iid}}{\sim} \Pi_0, \\
P_{1,n+1}, \dots, P_{1,n+m} &\stackrel{\text{iid}}{\sim} \Pi_1, \\
X_{i,1}, \dots, X_{i,K} \mid P_{0,i} &\stackrel{\text{iid}}{\sim} P_{0,i}, \quad i \in [n], \\
X_{n+i,j} \mid P_{0,n+i}, P_{1,n+i} &\stackrel{\text{indep}}{\sim} Q_{n+i,j}, \quad i \in [m], j \in [K],
\end{aligned}$$

where Π_0 and Π_1 are distributions over a measurable space of probability measures on $\mathcal{X}$. A test sample $X_{n+i,j}$ is classified as an *inlier* if it follows the null distribution conditionally (*i.e.*, $Q_{n+i,j} = P_{0,n+i}$). Otherwise (*i.e.*, $Q_{n+i,j} = P_{1,n+i}$), it is regarded as an *outlier*. Our objective is to identify groups containing outlying units or to detect outlying units directly within the test dataset.

These tasks can be framed as multiple testing problems. Group-level detection involves testing the following hypotheses:

$$H_i : Q_{n+i,j} = P_{0,n+i} \text{ for all } j \in [K], \quad (1)$$

for each $i \in [m]$. Let $\mathcal{H}_0 = \{i \in [m] : Q_{n+i,j} = P_{0,n+i} \text{ for all } j \in [K]\}$ be the index set of groups containing only inliers, and $\mathcal{H}_1 = [m] \setminus \mathcal{H}_0$. The proportions of null and non-null hypotheses are denoted by $\pi_0 = |\mathcal{H}_0|/m$ and $\pi_1 = 1 - \pi_0$, respectively.

For unit-level detection, we test the following hypotheses:

$$H_{i,j} : Q_{n+i,j} = P_{0,n+i}, \quad (2)$$

for each $i \in [m]$ and $j \in [K]$. With a slight abuse of notation, we let $\mathcal{H}_0 = \{(i,j) \in [m] \times [K] : Q_{n+i,j} = P_{0,n+i}\}$ be the index set of inliers, and $\mathcal{H}_1 = ([m] \times [K]) \setminus \mathcal{H}_0$ be the index set of outliers. In this case, the proportions of null and non-null hypotheses are denoted by $\pi_0 = |\mathcal{H}_0|/(mK)$ and $\pi_1 = 1 - \pi_0$, respectively.

A novelty detection procedure takes $\mathcal{D}_{\text{ref}}$ and $\mathcal{D}_{\text{test}}$ as inputs and produces a rejection set $\mathcal{R}$ of suspected outliers. The false discovery rate (FDR) is defined as the expectation of the false discovery proportion (FDP): $\text{FDR}(\mathcal{R}) = \mathbb{E}[\text{FDP}(\mathcal{R})]$, where

$$\text{FDP}(\mathcal{R}) = \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}. \quad (3)$$

Our goal is to develop procedures that control the FDR at a pre-specified level α .

2.2 Multiple testing with e-values

To design a valid novelty detection procedure with FDR control, we utilize e-values (Shafer, 2021; Vovk & Wang, 2021; Ramdas & Wang, 2025) and the eBH procedure (Wang & Ramdas, 2022). A test statistic e is called an *e-value* if it is non-negative and its expected value under the null hypothesis H_0 is at most one (i.e., $\mathbb{E}_{H_0}[e] \leq 1$). For a multiple testing problem with N hypotheses $H_1, \dots, H_N$ and corresponding e-values $e_1, \dots, e_N$, the eBH procedure rejects H_j if $e_j \geq N/(\alpha k^*)$, where $k^* = \max\{k \in [N] : e_{(k)} \geq N/(\alpha k)\}$ and $e_{(1)} \geq \dots \geq e_{(N)}$. We denote the resulting eBH rejection set by $\mathcal{R}_{\text{eBH}} = \{j \in [N] : e_j \geq N/(\alpha k^*)\}$. Wang & Ramdas (2022) demonstrated that this procedure controls the FDR at level α . Ignatiadis et al. (2024) further relaxed these conditions, showing that FDR is controlled if the non-negative statistics e_j satisfy the average expectation condition $\sum_{j \in \mathcal{H}_0} \mathbb{E}[e_j] \leq N$, where $\mathcal{H}_0 \subseteq [N]$ is the index set of the true nulls. These are referred to as *compound e-values*.

Lemma 2.1 (Ignatiadis et al. (2024)). *The eBH procedure based on valid compound e-values (i.e., $e_j \geq 0$ for all $j \in [N]$, and $\sum_{j \in \mathcal{H}_0} \mathbb{E}[e_j] \leq N$) at level α satisfies $\text{FDR}(\mathcal{R}_{\text{eBH}}) \leq \alpha$.*

We also consider the uniformly randomized eBH (U-eBH) procedure proposed by Xu & Ramdas (2025); see also Ramdas & Wang (2025). Given non-negative statistics $e_1, \dots, e_N$, U-eBH draws a single random variable $U \sim \text{Unif}(0, 1)$ independently of the data and applies the eBH procedure to the scaled statistics $e_1/U, \dots, e_N/U$. Let $k_U^* = \max\{k \in [N] : e_{(k)}/U \geq N/(\alpha k)\}$. We denote the resulting U-eBH rejection set by $\mathcal{R}_{\text{U-eBH}} = \{j \in [N] : e_j/U \geq N/(\alpha k_U^*)\}$. We note that U-eBH preserves the ordering of the statistics but randomizes the effective rejection threshold through the common multiplier $1/U$. Since $1/U \geq 1$, the U-eBH rejection set contains the eBH rejection set for every realization of U , while FDR control under arbitrary dependence is preserved as formalized in Lemma 2.2. In our numerical studies, U-eBH is used only as the final multiple-testing step and does not change the construction of the conformal e-values.

Lemma 2.2 (Xu & Ramdas (2025)). *The U-eBH procedure based on valid compound e-values (i.e., $e_j \geq 0$ for all $j \in [N]$, and $\sum_{j \in \mathcal{H}_0} \mathbb{E}[e_j] \leq N$) at level α satisfies $\text{FDR}(\mathcal{R}_{\text{U-eBH}}) \leq \alpha$.*

2.3 Hierarchical exchangeability

We clarify the specific form of exchangeability essential for constructing e-values. In manufacturing applications, two types of permutations typically preserve the data distribution: (1) permuting samples within the same group and (2) permuting the

Algorithm 1 Hierarchical Conformal Group-level Novelty Detection (HC-GND)

Input: Reference data $\mathcal{D}_{\text{ref}} = \{X_{i,j} : i \in [n], j \in [K]\}$, test data $\mathcal{D}_{\text{test}} = \{X_{n+i,j} : i \in [m], j \in [K]\}$, calibration index set $\mathcal{I}(\subseteq [n])$, target level α , threshold parameter $\tilde{\alpha}$.

- 1: Obtain calibration scores $\{S_i : i \in \mathcal{I}\}$ and test scores $\{S_{n+i} : i \in [m]\}$.
- 2: Compute the e-values e_i for $i \in [m]$ as defined in (4).
- 3: Sort the e-values in descending order: $e_{(1)} \geq e_{(2)} \geq \dots \geq e_{(m)}$.
- 4: Determine $k^* = \max\{k \in [m] : e_{(k)} \geq m/(\alpha k)\}$.

Output: Selection set $\mathcal{R} = \{i \in [m] : e_i \geq e_{(k^*)}\}$.

groups themselves. Notably, swapping samples across different groups generally alters the distribution. We formally define hierarchical exchangeability as follows:

Definition 2.1 (Hierarchical exchangeability). *The dataset $\{Z_i = (X_{i,1}, \dots, X_{i,K}) : i \in [n]\}$ is hierarchically exchangeable if:*

- (1) For any permutation $\sigma \in \mathcal{S}_n$,

$$(Z_1, \dots, Z_n) \stackrel{d}{=} (Z_{\sigma(1)}, \dots, Z_{\sigma(n)}).$$

- (2) For any $i \in [n]$ and $\tau \in \mathcal{S}_K$,

$$(Z_1, \dots, Z_n) \stackrel{d}{=} (Z_1, \dots, Z_{i-1}, (X_{i,\tau(1)}, \dots, X_{i,\tau(K)}), Z_{i+1}, \dots, Z_n).$$

This definition is more restrictive than those in Lee et al. (2023) and Lee & Ren (2025), which allow for variable group sizes. We focus on the fixed-group-size setting because it reflects a common feature of manufacturing data, where standardized product structures often induce repeated groups with the same number of units.

As shown in Section 2.4, this concept of exchangeability is sufficient for constructing conformal group-level novelty detection. Notably, under the null hypothesis (1), the augmented set $\{Z_1, \dots, Z_n, Z_{n+i}\}$ is hierarchically exchangeable. However, this property does not hold under the unit-level null hypothesis (2), as individual hypotheses do not guarantee the exchangeability of the corresponding test group. This makes it challenging to construct valid (compound) e-values without additional assumptions. To address this, we introduce assumptions on the nonconformity score function, as discussed in detail in Section 2.5.

2.4 Hierarchical conformal group-level novelty detection

Before constructing the e-values, we must specify a nonconformity score function. Fundamentally, hierarchical novelty detection evaluates how much an observation deviates

from its corresponding group distribution. We simplify this mechanism by measuring the deviation of an observation from the center of the group distribution. Let $C : \mathcal{X}^K \rightarrow \mathcal{X}$ denote the location functional that identifies the center of group observations, and let $s : \mathcal{X}^2 \rightarrow [0, \infty)$ be the nonconformity score function. Larger scores indicate greater nonconformity with the corresponding group distribution. Both the score function and the location functional may depend on the training data, although this dependence is not explicitly denoted here for brevity.

To construct the e-values, we first let $\mathcal{I} \subseteq [n]$ be the index set of calibration samples. We then define $S_{i,j} = s(X_{i,j}, C(Z_i))$ and $S_i = (S_{i,1}, \dots, S_{i,K})$ for $i \in \mathcal{I} \cup ([n+m] \setminus [n])$ and $j \in [K]$. Based on these scores, the e-values are defined as:

$$e_i = \frac{(|\mathcal{I}| + 1) \sum_{j=1}^K I(S_{n+i,j} \geq T_i)}{1 \vee \sum_{k \in \mathcal{I} \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j} \geq T_i)} \quad (4)$$

for each $i \in [m]$, where $T_i = \inf \{t \in \mathbb{R} : \text{FDP}_i(t) \leq \tilde{\alpha}\}$ is a stopping time based on the following FDP estimate,

$$\text{FDP}_i(t) = \frac{m-1}{|\mathcal{I}| + 1} \frac{\sum_{k \in \mathcal{I} \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j} \geq t)}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j} \geq t)}.$$

Here, $\tilde{\alpha}$ is a pre-specified threshold. Following Lee & Ren (2025), we use values slightly below α as practical choices; the sensitivity of the proposed procedures to $\tilde{\alpha}/\alpha$ is examined in Section 3.5. Finally, we apply the eBH procedure to identify the selection set. We refer to this procedure as *hierarchical conformal group-level novelty detection* (HC-GND), as summarized in Algorithm 1.

By leveraging the hierarchical exchangeability of the scores, we can show that the e-values defined in (4) are valid. The proof of the following theorem is provided in Appendix A.

Theorem 2.1. *For $i \in \mathcal{H}_0$, if $\{S_k : k \in \mathcal{I} \cup \{n+i\}\}$ is hierarchically exchangeable, then the e-value defined in (4) satisfies $\mathbb{E}[e_i] \leq 1$.*

There are several ways to construct scores that satisfy the hierarchical exchangeability required by Theorem 2.1. In this paper, we propose three procedures: the split, full, and group-wise methods.

Under the split conformal (SC) regime, the reference set is randomly partitioned into two disjoint subsets: a training set and a calibration set. Without loss of generality, we denote these as $\mathcal{D}_{\text{tr}} = \{X_{i,j} : i \in [n_{\text{tr}}], j \in [K]\}$ and $\mathcal{D}_{\text{cal}} = \{X_{n_{\text{tr}}+i,j} : i \in [n_{\text{cal}}], j \in [K]\}$, where $n = n_{\text{tr}} + n_{\text{cal}}$. In practice, we set $n_{\text{tr}} = n_{\text{cal}} = n/2$ by default. Using $\mathcal{D}_{\text{tr}}$, we construct the location functional $C(z) = C(z; \mathcal{D}_{\text{tr}})$ and the nonconformity score

function $s(x, x') = s(x, x'; \mathcal{D}_{\text{tr}})$. In this case, the calibration index set is $\mathcal{I} = [n] \setminus [n_{\text{tr}}]$. It is straightforward to verify that $\{S_k : k \in \mathcal{I} \cup \{n+i\}\}$ is hierarchically exchangeable for any $i \in \mathcal{H}_0$.

In the full conformal (FC) regime, we construct the location functional $C(z) = C(z; \mathcal{D})$ and the nonconformity score function $s(x, x') = s(x, x'; \mathcal{D})$ using the entire dataset $\mathcal{D} = \mathcal{D}_{\text{ref}} \cup \mathcal{D}_{\text{test}}$. We assume that this learning step is invariant under any group-wise permutation or within-group unit-wise permutation. The calibration index set is $\mathcal{I} = [n]$, and again, it can be verified that $\{S_k : k \in \mathcal{I} \cup \{n+i\}\}$ is hierarchically exchangeable for any $i \in \mathcal{H}_0$.

We note that the SC method may exhibit a loss of statistical power because it uses only a portion of the reference data for learning. In contrast, the FC method uses the reference and test sets jointly, potentially leading to a gain in statistical power by exploiting more data. However, there is a trade-off due to potential outlier contamination during the learning step. While the learning phase of SC exclusively utilizes $\mathcal{D}_{\text{tr}}$ (which is assumed to contain only inliers), the FC method inherently uses a potentially contaminated dataset. If training is conducted without accounting for these outliers, the resulting scores may fail to adequately distinguish outliers from inliers, leading to e-values with low expectations even for non-null hypotheses and thus diminishing the power of the procedure. To mitigate this, one strategy is to employ robust estimators. An alternative strategy is to reduce the impact of contamination by incorporating only specific blocks of the test set into the learning process, as proposed by Lee et al. (2026). In our hierarchical setting, the group structure provides a natural basis for this approach.

For each $i \in [m]$, we obtain the location functional $C^{(i)}(z) = C(z; \mathcal{D}_i)$ and the nonconformity score function $s^{(i)}(x, x') = s(x, x'; \mathcal{D}_i)$ using $\mathcal{D}_i = \mathcal{D}_{\text{ref}} \cup \{X_{n+i,j} : j \in [K]\}$. We then compute the scores $S_{k,j}^{(i)} = s^{(i)}(X_{k,j}, C^{(i)}(Z_k))$ for $k \in [n+m]$ and $j \in [K]$. As in the FC method, the calibration index set is $\mathcal{I} = [n]$. Based on these scores, the e-value for each $i \in [m]$ is computed as:

$$e_i = \frac{(|\mathcal{I}| + 1) \sum_{j=1}^K I(S_{n+i,j}^{(i)} \geq T_i)}{1 \vee \sum_{k \in \mathcal{I} \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j}^{(i)} \geq T_i)} \quad (5)$$

where T_i and $\text{FDP}_i(t)$ are defined analogously to the previous case but using the scores $\{S_k^{(i)} : k \in [n+m]\}$.

We refer to this method as hierarchical group-wise conformal group-level novelty detection (HGC-GND). This approach repeats the learning step m times, with each iteration incorporating only a single test group. Compared to the FC method, HGC-GND is less exposed to contamination from other test groups, allowing for the use

Algorithm 2 Hierarchical Split Conformal Unit-level Novelty Detection (HSC-UND)

Input: Reference data $\mathcal{D}_{\text{ref}} = \{X_{i,j} : i \in [n], j \in [K]\}$, test data $\mathcal{D}_{\text{test}} = \{X_{n+i,j} : i \in [m], j \in [K]\}$, target level α , threshold parameter $\tilde{\alpha}$.

- 1: Randomly partition the reference groups into two subsets: a training set $\mathcal{D}_{\text{tr}} = \{X_{i,j} : i \in [n_{\text{tr}}], j \in [K]\}$ and a calibration set $\mathcal{D}_{\text{cal}} = \{X_{n_{\text{tr}}+i,j} : i \in [n_{\text{cal}}], j \in [K]\}$, where $n_{\text{tr}} + n_{\text{cal}} = n$.
- 2: Construct the location functional $C : \mathcal{X}^K \rightarrow \mathcal{X}$ and the nonconformity score function $s : \mathcal{X}^2 \rightarrow [0, \infty)$ using $\mathcal{D}_{\text{tr}}$.
- 3: Compute the scores $S_{i,j} = s(X_{i,j}, C(Z_i))$ for all $i \in [n+m] \setminus [n_{\text{tr}}]$ and $j \in [K]$.
- 4: Compute the e-values $e_{i,j}$ for $i \in [m], j \in [K]$ as defined in (6).
- 5: Sort the e-values in descending order: $e_{(1)} \geq e_{(2)} \geq \dots \geq e_{(N_{\text{test}})}$, where $N_{\text{test}} = mK$.
- 6: Determine $k^* = \max\{k \in [N_{\text{test}}] : e_{(k)} \geq N_{\text{test}}/(\alpha k)\}$.

Output: Selection set $\mathcal{R} = \{(i, j) \in [m] \times [K] : e_{i,j} \geq e_{(k^*)}\}$.

of standard methods without requiring specialized robust corrections. For $i \in \mathcal{H}_0$, the validity of the e-value relies on the hierarchical exchangeability of $\{S_k^{(i)} : k \in \mathcal{I} \cup \{n+i\}\}$, which leads to the following corollary.

Corollary 2.1. *For $i \in \mathcal{H}_0$, the e-value defined in (5) satisfies $\mathbb{E}[e_i] \leq 1$.*

2.5 Hierarchical conformal unit-level novelty detection

We begin by introducing our procedure. Similar to the split method for group-level novelty detection, we obtain the nonconformity scores $\{S_{k,j} : k \in \mathcal{I} \cup ([n+m] \setminus [n]), j \in [K]\}$, where $\mathcal{I} = [n] \setminus [n_{\text{tr}}]$. Using these scores, we define the e-values as:

$$e_{i,j} = \frac{K(n_{\text{cal}} + 1)I(S_{n+i,j} \geq T_i^+)}{\sum_{k \in \mathcal{I} \cup \{n+i\}} \sum_{l=1}^K I(S_{k,l} \geq T_i^-) + \epsilon}, \quad (6)$$

where $\epsilon, T_{\min} > 0$ are predetermined constants, and $T_i^+ = \inf \{t \in [T_{\min}, \infty) : \text{FDP}_i^+(t) \leq \tilde{\alpha}\}$ and $T_i^- = \inf \{t \in [T_{\min}, \infty) : \text{FDP}_i^-(t) \leq \tilde{\alpha}\}$ are stopping times based on the following FDP estimates,

$$\begin{aligned} \text{FDP}_i^+(t) &= \frac{m-1}{n_{\text{cal}}+1} \frac{\sum_{k \in \mathcal{I} \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j} \geq t) + \epsilon}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j} \geq t)}, \\ \text{FDP}_i^-(t) &= \frac{m-1}{n_{\text{cal}}+1} \frac{\sum_{k \in \mathcal{I}} \sum_{j=1}^K I(S_{k,j} \geq t)}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j} \geq t)}. \end{aligned}$$

Finally, we apply the eBH procedure to identify the selection set. We refer to this procedure as *hierarchical split conformal unit-level novelty detection* (HSC-UND), as summarized in Algorithm 2.

As discussed in Section 2.3, the individual hypotheses (2) in unit-level novelty detection do not inherently guarantee the hierarchical exchangeability of the augmented set $\{Z_1, \dots, Z_n, Z_{n+i}\}$. To facilitate this, we consider a “null” version of Z_{n+i} , denoted by Z_{n+i}^* . Specifically, for an inlier index $(i, j) \in \mathcal{H}_0$, $X_{n+i,j}^*$ is identical to the original $X_{n+i,j}$. For an outlier index $(i, j) \in \mathcal{H}_1$, we assume that $X_{n+i,j}^* \mid P_{0,n+i}, P_{1,n+i} \sim P_{0,n+i}$. Under this construction, the augmented set $\{Z_1, \dots, Z_n, Z_{n+i}^*\}$ is hierarchically exchangeable. Since we only modify the test group observations while maintaining the SC regime, the corresponding scores remain unchanged except for those in the $(n+i)$ -th group. Let $S_{n+i,j}^*$ for $j \in [K]$ denote the null version of the $(n+i)$ -th group scores. We then define the null version of the e-values as

$$e_{i,j}^* = \frac{K(n_{\text{cal}} + 1)I(S_{n+i,j}^* \geq T_i^*)}{1 \vee \left\{ \sum_{k \in \mathcal{I}} \sum_{l=1}^K I(S_{k,l} \geq T_i^*) + \sum_{l=1}^K I(S_{n+i,l}^* \geq T_i^*) \right\}}, \quad (7)$$

where $T_i^* = \inf \{t \in [T_{\min}, \infty) : \text{FDP}_i^*(t) \leq \tilde{\alpha}\}$ is a stopping time based on the FDP estimate,

$$\text{FDP}_i^*(t) = \frac{m-1}{n_{\text{cal}} + 1} \frac{\sum_{k \in \mathcal{I}} \sum_{j=1}^K I(S_{k,j} \geq t) + \sum_{j=1}^K I(S_{n+i,j}^* \geq t)}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j} \geq t)}.$$

The hierarchical exchangeability of $\{Z_k : k \in \mathcal{I}\} \cup \{Z_{n+i}^*\}$ implies the hierarchical exchangeability of the scores $\{S_k : k \in \mathcal{I}\} \cup \{S_{n+i}^*\}$. Hence, the null e-values in (7) satisfy $\mathbb{E}[e_{i,j}^*] \leq 1$. Although these e-values are not practically computable as they rely on the hypothetical Z_{n+i}^* , they serve to establish the validity of the e-values in (6) under the condition stated below.

Theorem 2.2. *For $i \in [m]$ and $j \in [K]$, define $X_{n+i,j}^* = X_{n+i,j}$ if $Q_{n+i,j} = P_{0,n+i}$, and $X_{n+i,j}^* \mid P_{0,n+i}, P_{1,n+i} \sim P_{0,n+i}$ independent of $X_{n+i,j}$ if $Q_{n+i,j} = P_{1,n+i}$. Let $Z_{n+i}^* = (X_{n+i,1}^*, \dots, X_{n+i,K}^*)$ and $S_{n+i,j}^* = s(X_{n+i,j}^*, C(Z_{n+i}^*))$. Assume there exist positive constants $\epsilon, T_{\min} > 0$ such that*

$$\sum_{j:(i,j) \in \mathcal{H}_0} I(S_{n+i,j} \geq t) \leq \sum_{j=1}^K I(S_{n+i,j}^* \geq t) \leq \sum_{j=1}^K I(S_{n+i,j} \geq t) + \epsilon \quad (8)$$

for all $t \in [T_{\min}, \infty)$ almost surely. Then, the e-values defined in (6) are valid compound e-values. That is, $\sum_{(i,j) \in \mathcal{H}_0} \mathbb{E}[e_{i,j}] \leq mK$.

Condition (8) provides the necessary flexibility for the e-values to satisfy the average expectation condition, ensuring their validity as compound e-values. Compared to standard e-value constructions, the inclusion of these constants may reduce

statistical power, and thus, they must be chosen judiciously. While adaptive data-driven selection is possible, the constants must be determined solely from the training set to preserve exchangeability. By default, we propose $\epsilon = 0.1K$ and $T_{\min} = Q_{1-\gamma}((n_{\text{tr}}K)^{-1} \sum_{i=1}^{n_{\text{tr}}} \sum_{j=1}^K \delta_{S_{i,j}})$ with $\gamma = 2\alpha$, where $Q_{\gamma}(P) = \sup\{t \in \mathbb{R} : P_{X \sim P}(X \leq t) \leq \gamma\}$.

One might also consider applying the FC or group-wise regimes to the UND problem (see Appendix B). However, establishing the theoretical validity of these approaches is challenging. In the FC and GC regimes, replacing non-null observations with hypothetical nulls affects the entire set of scores-including those of the reference and other test groups. This complex interdependence makes it difficult to analyze the properties of e-values under these regimes. Despite these theoretical difficulties, we evaluate their empirical performance in Section 3.

3 Numerical study

In this section, we numerically compare the proposed methods with existing methods that ignore the hierarchical structure for both group-level and unit-level novelty detection problems (Sections 3.3 and 3.4). We also examine how the choice of $\tilde{\alpha}/\alpha$ affects the performance of the proposed and existing methods (Section 3.5).

3.1 Settings

We partially adapt the outlier-generation settings from Bates et al. (2023), Jin & Candès (2023), Lee & Ren (2025), and Lee et al. (2026). The numbers of groups in the reference and test datasets are set to $n = |\mathcal{D}_{\text{ref}}| = 50$ and $m = |\mathcal{D}_{\text{test}}| = 150$, respectively. Each group consists of $K = 10$ units, and the dimension of the measurements for each unit is $d = 30$. We generate $W_i \sim \text{Unif}([-3, 3]^d)$ for $i \in [n + m]$ and $L_{i,j} \sim N_d(0, I_d)$ for $i \in [n + m]$ and $j \in [K]$. The reference data are generated as $X_{i,j} = W_i + L_{i,j}$ for $i \in [n]$ and $j \in [K]$. For the test data, we consider different outlier-generation mechanisms for the GND and UND problems.

For the GND problem, we randomly select $m\pi_1$ groups from $[m]$, define them as $\mathcal{H}_1$, and set $\mathcal{H}_0 = [m] \setminus \mathcal{H}_1$. For each $i \in \mathcal{H}_1$, we generate the number K_i of anomalous units from the following probability distribution:

$$P(K_i = 1) = 0.5, \quad P(K_i = 2) = 0.3, \quad P(K_i = 3) = 0.2.$$

Then, we set

$$X_{n+i,j} = \begin{cases} W_{n+i} + \sqrt{a}L_{n+i,j}, & \text{if } i \in \mathcal{H}_1, j \in [K_i], \\ W_{n+i} + L_{n+i,j}, & \text{otherwise.} \end{cases}$$

For the UND problem, we randomly select $mK\pi_1$ elements from $[m] \times [K]$, define them as $\mathcal{H}_1$, and set $\mathcal{H}_0 = [m] \times [K] \setminus \mathcal{H}_1$. Then, we set

$$X_{n+i,j} = \begin{cases} W_{n+i} + \sqrt{a}L_{n+i,j}, & \text{if } (i,j) \in \mathcal{H}_1, \\ W_{n+i} + L_{n+i,j}, & \text{otherwise.} \end{cases}$$

The proportion of outliers in the test set, denoted by π_1 , ranges from 0.1 to 0.5, and the signal strength a ranges from 1 to 4.5. The target FDR level is set to $\alpha = 0.1$, the threshold parameter is set to $\tilde{\alpha} = 0.9\alpha$, and ϵ used in the definition of the HC-UND e-value in (6) is set to 1. All experiments are repeated 1000 times.

3.2 Nonconformity score function and location functional

As discussed in Section 2.4, we need to specify the nonconformity score function $s : \mathcal{X}^2 \rightarrow [0, \infty)$ and the location functional $C : \mathcal{X}^K \rightarrow \mathcal{X}$. The location functional C should be permutation invariant. Although our framework does not impose explicit requirements on these functions, the assumption in (8) implicitly requires certain desirable properties.

To illustrate this point, suppose that a group contains some outliers. Let $S_1, \dots, S_K$ denote the observed scores, and let $S_1^*, \dots, S_K^*$ denote their null-version scores. The outlying units are expected to have relatively large scores compared with the inlying units. At the same time, the scores of the inlying units should not differ substantially from their null-version scores. Ideally, if the scores of the inlying units are identical to their null-version scores, then the assumption in (8) holds. In contrast, if these differences are nonnegligible, then the assumption in (8) may fail to hold. Therefore, we need a nonconformity score function and a location functional that do not substantially change the scores of the inlying units.

We propose to use a multivariate version of the trimmed mean, defined by

$$C(z; \Sigma) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} x_j, \quad (9)$$

where $z = (x_1, \dots, x_K) \in \mathcal{X}^K$, Σ is the within-group covariance matrix,

$$\hat{f}_j = \sum_{k=1}^K \exp \left(-\frac{1}{2} (x_j - x_k)^\top \Sigma^{-1} (x_j - x_k) \right),$$

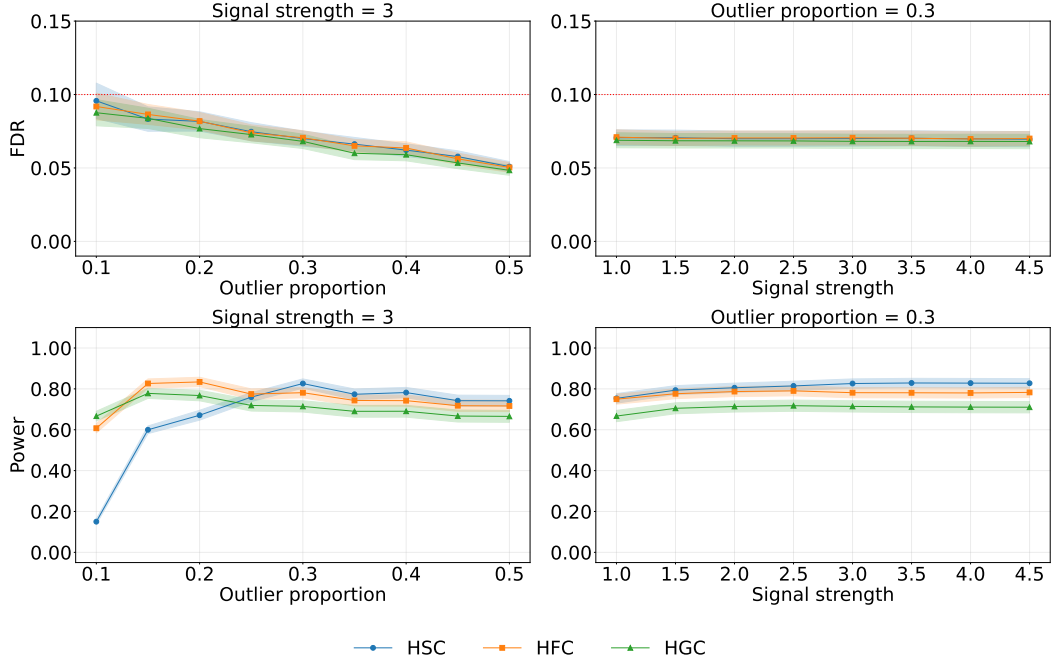


Figure 1: Empirical FDR and power of the proposed HC-GND procedures under varying outlier proportions and signal strengths. The top panels report the empirical FDR, and the bottom panels report the empirical power. The left panels vary the outlier proportion with the signal strength fixed at $a = 3$, whereas the right panels vary the signal strength with the outlier proportion fixed at $\pi_1 = 0.3$. The horizontal red dotted line in the FDR plots indicates the target FDR level $\alpha = 0.1$.

and $\mathcal{J} = \{j : \hat{f}_j \geq Q_\beta \left(K^{-1} \sum_{k=1}^K \delta_{\hat{f}_k} \right)\}$. Here, the quantities $\hat{f}_j$ can be viewed as kernel density estimates, where the within-group covariance matrix is used as the bandwidth matrix. The set $\mathcal{J}$ consists of units whose kernel density estimates exceed the β -quantile. In practice, we use $\beta = 0.5$. Thus, $C(z; \Sigma)$ averages the units with relatively high kernel density estimates, which are regarded as inliers. We then use the squared Mahalanobis distance as the nonconformity score, defined by $s(x, x'; \Sigma) = (x - x')^\top \Sigma^{-1} (x - x')$. Outliers are expected to have small kernel density estimates, so they are unlikely to be included in $\mathcal{J}$. Consequently, replacing outliers with inliers has only a limited effect on the location estimate, which is the desired property. The within-group covariance matrix Σ is estimated using the training set $\mathcal{D}_{\text{tr}}$ under the SC regime, the entire dataset $\mathcal{D} = \mathcal{D}_{\text{ref}} \cup \mathcal{D}_{\text{test}}$ under the FC regime, and $\mathcal{D}_i = \mathcal{D}_{\text{ref}} \cup \{X_{n+i,j} : j \in [K]\}$ under the GC regime.

3.3 Group-level novelty detection

In the group-level novelty detection problem, there are no directly comparable existing methods; thus, we compare the three proposed HC-GND procedures: HSC, HFC, and HGC. Figure 1 reports the empirical FDR and power of the proposed HC-GND

procedures under varying outlier proportions and signal strengths. Overall, the empirical FDR of all three proposed methods remains below the target level across the considered settings. This indicates that the proposed HC-GND procedures maintain empirical FDR control even when the outlier proportion or signal strength changes.

The power results reveal clear differences among the three methods. When the signal strength is fixed at $a = 3$, HFC and HGC maintain relatively high power as the outlier proportion decreases, whereas the power of HSC decreases sharply. This behavior may be attributed to the fact that HFC and HGC use larger datasets for score calibration than HSC. When the outlier proportion is small, contamination from anomalous groups is limited; therefore, the larger calibration datasets can improve the stability of score calibration without being substantially affected by outliers. Consequently, HFC and HGC retain high power even in sparse-outlier settings. When the outlier proportion is fixed, the power of all three methods remains stable or slightly increases as the signal strength increases.

3.4 Unit-level novelty detection

For the unit-level novelty detection problem, we compare the proposed procedures with the methods of Lee et al. (2026), which do not account for the hierarchical structure. Specifically, we consider SC, FC, and AD as benchmark methods. AD is AdaDetect, which is based on the BH procedure applied to conformal p -values, and SC and FC are based on the U-eBH procedure applied to conformal e -values. Following Lee et al. (2026), FC uses Isolation Forest as the working model, and SC uses one-class SVM.

Figure 2 reports the empirical FDR and power for unit-level novelty detection under varying outlier proportions and signal strengths. Overall, the empirical FDR of the proposed hierarchical procedures remains below the target level across the considered settings. This shows that incorporating the hierarchical structure does not compromise empirical FDR control in the numerical experiments. Although the theoretical FDR control of HFC and HGC has not been established, the simulation results show that both methods empirically control the FDR across a wide range of settings. Moreover, HFC and HGC generally achieve higher power than HSC. The power results show that the proposed hierarchical methods generally outperform the non-hierarchical methods of Lee et al. (2026). This improvement can be explained by the fact that unit-level anomalies are defined relative to the hierarchical structure. Some units may be distinguishable as outliers within their own groups, but may not appear sufficiently anomalous when the hierarchical structure is ignored. By accounting for within-group dependence and group-specific structure, the proposed methods can better identify such anomalous units. Consequently, the proposed HC-UND procedures

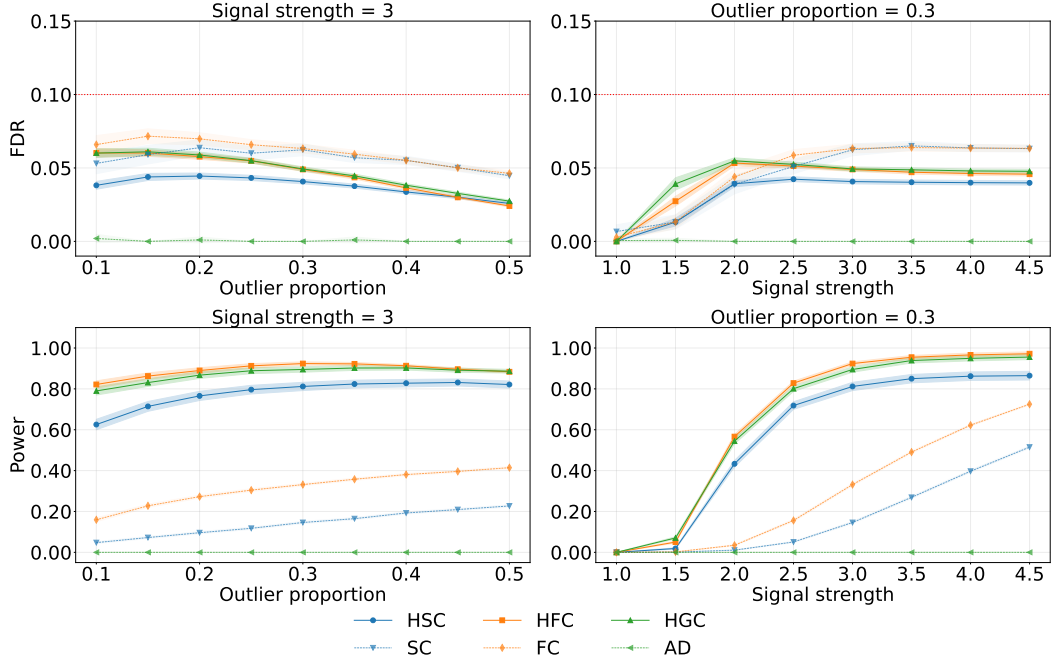


Figure 2: Empirical FDR and power of the UND procedures under varying outlier proportions and signal strengths. The proposed methods (HSC, HFC, and HGC) are compared with non-hierarchical methods (SC, FC, and AD). The top panels report the empirical FDR, and the bottom panels report the empirical power. The left panels vary the outlier proportion with the signal strength fixed at $a = 3$, whereas the right panels vary the signal strength with the outlier proportion fixed at $\pi_1 = 0.3$. The horizontal red dotted line in the FDR plots indicates the target FDR level $\alpha = 0.1$.

achieve higher empirical power while maintaining empirical FDR control.

3.5 Sensitivity analysis for selection of $\tilde{\alpha}/\alpha$

The tuning parameter $\tilde{\alpha}$ determines the stopping threshold used to construct the conformal e-values in both HC-GND (Section 2.4) and HC-UND (Section 2.5). Although $\tilde{\alpha}$ is not the target FDR level itself, it affects the rejection behavior through the data-dependent threshold at which the FDP estimate first falls below $\tilde{\alpha}$. The effect of $\tilde{\alpha}$ is not necessarily monotone because it changes both the score threshold and the resulting e-values. A larger value of $\tilde{\alpha}$ relaxes the stopping condition, so the stopping time can be reached at a smaller score threshold. Lowering the score threshold increases the number of scores exceeding the threshold, but these additional exceedances enter both the numerator and denominator of the e-value construction. Thus, depending on whether the denominator grows faster than the numerator, increasing $\tilde{\alpha}$ may either increase or decrease the final rejection power. Prior work suggests that choosing $\tilde{\alpha}$ below α can improve power; for example, Ren & Barber (2024) reported strong empirical performance with $\tilde{\alpha} = 0.5\alpha$, while Lee & Ren (2025) used $\tilde{\alpha} = 0.9\alpha$ as a practical default.

Motivated by these findings, we examine the sensitivity of the proposed procedures to the ratio $\tilde{\alpha}/\alpha$.

Figures 6, 7, 8, and 9 summarize the simulation results for HC-GND and HC-UND under split and full conformal regimes, as reported in Appendix C. Across the considered values of $\tilde{\alpha}/\alpha$, the empirical FDR remains stable around or below the target level, and power increases as either the signal strength or the outlier proportion increases. The variation induced by $\tilde{\alpha}/\alpha$ is comparatively modest relative to the effects of signal strength and outlier proportion, suggesting that the proposed procedures are not highly sensitive to this tuning parameter in the simulated settings.

Based on the weak sensitivity to $\tilde{\alpha}/\alpha$ observed in these simulations, we set $\tilde{\alpha} = 0.9\alpha$ throughout the numerical experiments. This choice follows the practical recommendation of Lee & Ren (2025), avoids additional data-driven tuning on the test set, and yields stable empirical performance in the sensitivity analysis.

4 Novelty detection in EV battery pack dataset

4.1 EV battery pack data

We analyze the publicly available EV battery-pack manufacturing dataset provided by KAMP (Ministry of SMEs and Startups & KAIST, 2022). The original data include both charging and discharging tests, but the present analysis uses only charging voltage trajectories because voltage changes during charging provide direct information about cell-level behavior. After excluding packs with incomplete charging trajectories, the normal reference pool contains 39 battery packs. The original KAMP data consist of this reference set and a test set with four packs, including two normal packs and two packs labeled abnormal. Because the original test set is small, we randomly select two packs from the normal reference pool and treat them as additional test packs. Consequently, the reference set used in this analysis contains 37 normal battery packs, whereas the test set contains six battery packs: four normal packs and two abnormal packs. Throughout this analysis, Normal packs 0 and 1 and Abnormal packs 4 and 5 denote the original KAMP test packs, whereas Normal packs 2 and 3 denote the additional packs selected from the normal reference pool. The abnormality labels are available only at the pack level, and no module- or cell-level ground truth is provided.

Each battery pack consists of 16 modules, and each module contains 11 cells. We therefore translate the battery data into the two-level hierarchy used by the proposed procedures by treating each module as a group and each cell within a module as a unit, while pack-level dependence among modules is not modeled explicitly. Thus, in

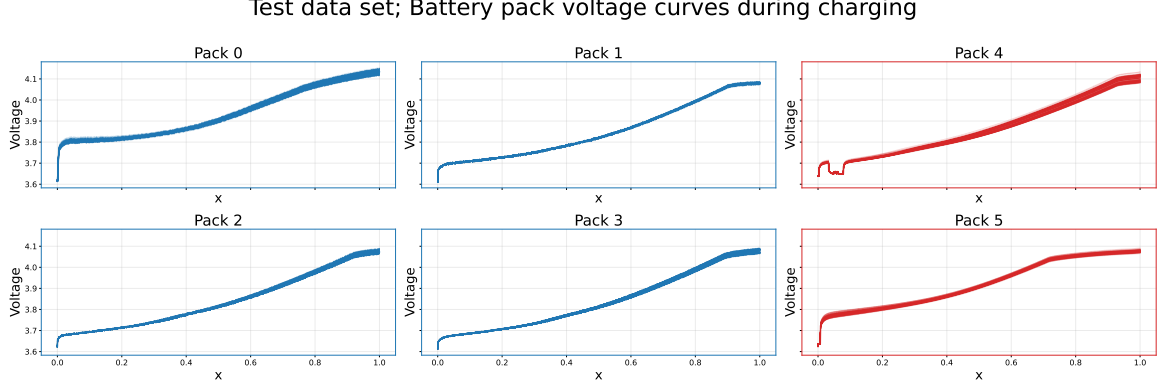


Figure 3: Preprocessed charging voltage curves for all cells in the six test packs used in this analysis. Normal packs 0 and 1 and Abnormal packs 4 and 5 come from the original KAMP test set, whereas Normal packs 2 and 3 are randomly selected from the normal reference pool and treated as additional test packs.

the notation of Section 2.1, the battery-pack application has $n = 37 \times 16 = 592$, $m = 6 \times 16 = 96$, and $K = 11$.

The charging voltage time series of each cell is treated as a functional observation. Since charging durations differ across battery packs, we extract the charging interval for each cell and normalize its time domain. Let $V_{ij} : [0, t_{ij}] \rightarrow \mathbb{R}$ denote the charging voltage curve of the j -th cell in the i -th module. To place all charging curves on a common time domain, we define the time-normalized curve $\tilde{V}_{ij} : [0, 1] \rightarrow \mathbb{R}$ by $\tilde{V}_{ij}(s) = V_{ij}(st_{ij})$ for $s \in [0, 1]$. Figure 3 displays the resulting preprocessed charging curves for all cells in the test packs.

We then use functional principal component analysis (FPCA) (Lei et al., 2015) as a feature extraction step. For each conformal regime, the FPCA representation is fitted using the data available in that regime, and both reference and test charging curves are projected onto the first two principal components. We denote the resulting two-dimensional score vector by X_{ij} for each cell. In the split conformal regime, FPCA is fitted using the training subset of the reference set; in the full conformal regime, it is fitted using the combined reference and test sets; and in the group-wise conformal regime, it is fitted using the reference set together with the corresponding test group. Figure 4 visualizes these two-dimensional FPCA score vectors under the split conformal construction.

After this transformation, all novelty detection procedures are applied to the two-dimensional PC score vectors X_{ij} . For the location functional C and the nonconformity score function s , we use the same trimmed-mean location functional and squared Mahalanobis distance introduced in Section 3.2, applied to the FPCA score vectors. Specifically, for each module, we compute the location functional C from the within-

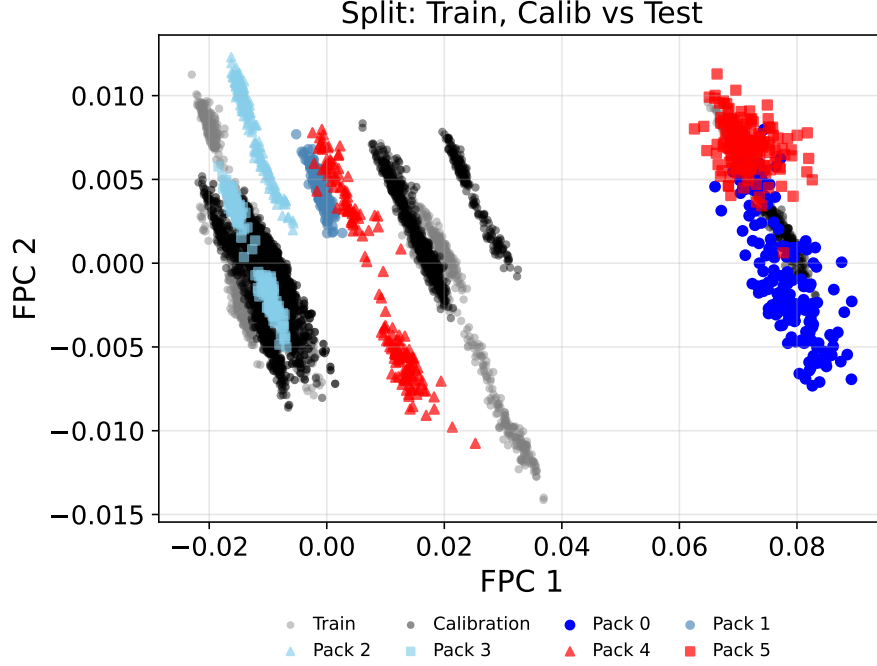


Figure 4: Scatter plot of the two-dimensional FPCA score vectors used for novelty detection under the split conformal construction. Each point represents a cell-level observation after projecting its charging curve onto the first two principal components.

module vectors and use the nonconformity score $S_{ij} = s(X_{ij}, C(Z_i))$, where $Z_i = (X_{i1}, \dots, X_{iK})$. These scores are then used to construct the HC-GND and HC-UND e-values as described in Sections 2.4 and 2.5. For comparison, we also apply the non-hierarchical SC, FC, and AdaDetect novelty detection methods to the cell-level PC score vectors, treating cells as ungrouped observations and ignoring the module–cell hierarchy.

4.2 Detection results for the EV battery data

We next examine how the feature patterns in Figures 3 and 4 translate into novelty detection results. Figure 5 provides a combined pack-wise comparison of the hierarchical procedures and non-hierarchical baselines. The first three columns show HC-GND and HC-UND rejections under the split, full, and group-wise conformal regimes, whereas the last three columns show non-hierarchical cell-level detections obtained by SVM, Isolation Forest, and AdaDetect while ignoring the module–cell hierarchy.

Figure 4 shows that Normal pack 0 is more dispersed in the FPCA score space than the other normal packs, with several cells located away from the main normal cluster. This pattern is consistent with the charging curves in Figure 3, where Normal pack 0 exhibits greater cell-to-cell variability during charging. Although this pack is

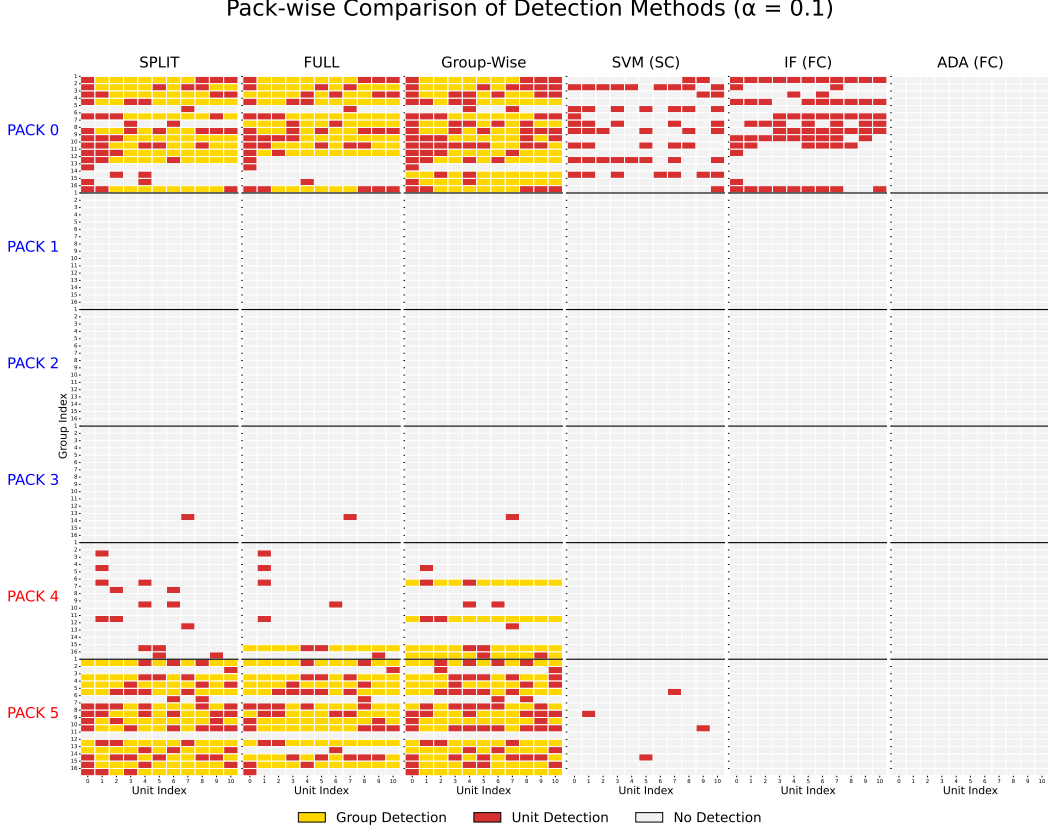


Figure 5: Combined pack-wise comparison of hierarchical and non-hierarchical novelty detection results for the EV battery-pack data at $\alpha = 0.1$. Left three columns: hierarchical split, full, and group-wise HC-GND/HC-UND results; yellow marks HC-GND module rejections and red marks HC-UND cell rejections. Right three columns: non-hierarchical SVM, IF, and AdaDetect results; red marks rejected cells and gray marks no detection.

labeled normal at the pack level, the feature representation suggests that it may contain localized charging patterns that differ from the normal reference behavior.

The hierarchical columns in Figure 5 show that HSC, HFC, and HGC detect anomalous signals in Normal pack 0 despite its normal pack-level label, while the remaining normal packs show few or no rejections. The two abnormal packs display different patterns: Abnormal pack 4 has only limited detections, whereas Abnormal pack 5 shows extensive rejections across multiple modules and cells. These results indicate that pack-level labels do not fully describe the localized module- and cell-level structure revealed by the proposed procedures.

The same figure also compares the hierarchical results with the non-hierarchical baseline methods of Lee et al. (2026). SVM and Isolation Forest produce many cell-level rejections in Normal pack 0, but they detect little or no structure in the abnormal packs, and AdaDetect produces no rejections in this analysis. By contrast, the proposed hierarchical procedures aggregate evidence within modules and identify module-level

and cell-level signals in Abnormal pack 5 that are largely missed when the hierarchy is ignored. This comparison suggests that incorporating the module–cell hierarchy provides useful diagnostic information for battery-pack anomaly detection when only pack-level labels are available.

5 Discussion

This paper proposes a conformal e-value framework for unsupervised novelty detection in hierarchically structured data. The main goal is to detect anomalous groups and anomalous units while controlling the FDR, without reducing the data to an unstructured collection of observations. The proposed HC-GND and HC-UND procedures use within-group nonconformity scores to preserve the group structure, and then use e-value-based multiple testing procedures to convert these scores into rejection sets. This separation between score construction and error control makes the framework flexible: different feature representations or score functions can be used, as long as the relevant exchangeability conditions are respected.

The theoretical results also show that group-level and unit-level hierarchical novelty detection have different statistical difficulties. For group-level detection, the null hypothesis implies hierarchical exchangeability of the augmented reference and test groups, so valid conformal e-values can be obtained under the split, full, and group-wise conformal regimes. For unit-level detection, however, a single null unit does not make the entire test group exchangeable. The validity result for HSC-UND therefore relies on a comparison between the observed test group and a hypothetical null version of that group, leading to a compound e-value condition. This distinction explains why HFC-UND and HGC-UND are useful empirical extensions, but their theoretical FDR guarantees require further analysis.

The numerical studies support the practical value of incorporating the hierarchy. In the group-level simulations, the proposed methods maintain empirical FDR control, and the full and group-wise regimes often gain power by using more data for score construction. In the unit-level simulations, the hierarchical methods achieve stable empirical FDR and generally outperform non-hierarchical alternatives, especially when anomalous units are more visible relative to their own groups than relative to the pooled population. The sensitivity analysis further suggests that the proposed procedures are not highly sensitive to the choice of $\tilde{\alpha}/\alpha$ in the simulated settings, which supports the use of $\tilde{\alpha} = 0.9\alpha$ in the experiments.

The EV battery-pack analysis illustrates how these methodological differences appear in a real manufacturing dataset. The proposed procedures identify localized

module- and cell-level irregularities that are not directly encoded in the pack-level KAMP labels. In particular, the analysis finds substantial rejections in Normal pack 0 and Abnormal pack 5, while the non-hierarchical baselines either concentrate mostly on Normal pack 0 or produce no rejection in the AdaDetect case. These results should not be interpreted as replacing engineering diagnosis, because module- and cell-level ground truth is unavailable. Rather, they show that hierarchical conformal novelty detection can provide an additional diagnostic layer when only coarse pack-level labels are observed.

Several limitations remain. The framework assumes a fixed group size, which is appropriate for many standardized manufacturing systems but restrictive for domains with variable group sizes. The EV analysis also treats modules as groups and does not explicitly model dependence among modules within the same pack. In addition, the data analysis uses only charging voltage curves and a two-dimensional FPCA representation, so richer functional features or multimodal measurements may improve detection in future work. Further theoretical work is needed for unit-level full and group-wise conformal procedures, and for adaptive choices of score functions, tuning parameters, and model-selection strategies that preserve the exchangeability structure. Overall, the results suggest that conformal e-values provide a useful foundation for FDR-controlled novelty detection in hierarchical manufacturing data, while leaving several important directions for broader theory and more detailed industrial validation.

References

- Bae, J., Lee, J.-H., & Kim, S. (2023). Pni: industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6373–6383).
- Bates, S., Candès, E., Lei, L., Romano, Y., & Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1), 149–178.
- Bertolini, M., Mezzogori, D., Neroni, M., & Zammori, F. (2021). Machine learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, 175, 114820.
- Bolton, R. J. & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, 17(3), 235–255.
- Fernando, T., Gammulle, H., Denman, S., Sridharan, S., & Fookes, C. (2021). Deep

- learning for medical anomaly detection—a survey. *ACM Computing Surveys (CSUR)*, 54(7), 1–37.
- Guo, H., Yuan, S., & Wu, X. (2021). Logbert: Log anomaly detection via bert. In *2021 international joint conference on neural networks (IJCNN)* (pp. 1–8).: IEEE.
- He, S., Zhu, J., He, P., & Lyu, M. R. (2016). Experience report: System log analysis for anomaly detection. In *2016 IEEE 27th international symposium on software reliability engineering (ISSRE)* (pp. 207–218).: IEEE.
- Ignatiadis, N., Wang, R., & Ramdas, A. (2024). Asymptotic and compound e-values: multiple testing and empirical bayes. *arXiv preprint arXiv:2409.19812*.
- Jin, Y. & Candès, E. J. (2023). Model-free selective inference under covariate shift via weighted conformal p-values. *arXiv preprint arXiv:2307.09291*.
- Kim, D., Kang, P., Cho, S., Lee, H.-j., & Doh, S. (2012). Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. *Expert Systems with Applications*, 39(4), 4075–4083.
- Lee, J., Popov, I., & Ren, Z. (2026). Full-conformal novelty detection: A powerful and non-random approach. *arXiv preprint arXiv:2501.02703*.
- Lee, Y., Barber, R. F., & Willett, R. (2023). Distribution-free inference with hierarchical data. *arXiv preprint arXiv:2306.06342*.
- Lee, Y. & Ren, Z. (2025). Selection from hierarchical data with conformal e-values. *arXiv preprint arXiv:2501.02514*.
- Lei, J., Rinaldo, A., & Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1), 29–43.
- Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., & Jin, Y. (2024). Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1), 104–135.
- Ministry of SMEs and Startups & KAIST (2022). Electronic components (battery pack) quality assurance ai dataset. Korea AI Manufacturing Platform (KAMP).
- Ramdas, A. & Wang, R. (2025). Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2), 1–390.

- Ren, Z. & Barber, R. F. (2024). Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.
- Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2), 407–431.
- Vovk, V. & Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3), 1736–1754.
- Wang, R. & Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3), 822–852.
- Xu, Z. & Ramdas, A. (2025). More powerful multiple testing under dependence via randomization. *arXiv preprint arXiv:2305.11126*.

A Proofs

A.1 Proof of Theorem 2.1

Proof. By construction, $\text{FDP}_i(t)$ is invariant under the permutation of $\{S_k : k \in \mathcal{I}\} \cup \{S_{n+i}\}$ for any $t \in \mathbb{R}$, and consequently, the same invariance holds for T_i . Using this property, we observe the following,

$$\frac{1}{|\mathcal{I}| + 1} \mathbb{E}[e_i] = \mathbb{E} \left[\frac{\sum_{j=1}^K I(S_{n+i,j} \geq T_i)}{1 \vee \sum_{k \in \mathcal{I} \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j} \geq T_i)} \right] \leq \frac{1}{|\mathcal{I}| + 1},$$

where the last inequality is due to the hierarchical exchangeability of $\{S_k : k \in \mathcal{I}\} \cup \{S_{n+i}\}$. This completes the proof. $\square$

A.2 Proof of Theorem 2.2

Proof. We first introduce an “oracle” FDP estimate, $\text{FDP}_i^*(t)$, defined as

$$\text{FDP}_i^*(t) = \frac{m-1}{n_{\text{cal}} + 1} \cdot \frac{\sum_{k \in \mathcal{I}} \sum_{j=1}^K I(S_{k,j} \geq t) + \sum_{j=1}^K I(S_{n+i,j}^* \geq t)}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j} \geq t)},$$

with its corresponding stopping time

$$T_i^* = \inf \{t \in [T_{\min}, \infty) : \text{FDP}_i^*(t) \leq \tilde{\alpha}\}.$$

By construction, $\text{FDP}_i^*(t)$ is invariant under permutations of $\{Z_k : k \in \mathcal{I}\} \cup \{Z_{n+i}^*\}$ for any $t \in [T_{\min}, \infty)$, and consequently, this invariance also holds for T_i^* . Under assumption (8), it follows that $\text{FDP}_i^-(t) \leq \text{FDP}_i^*(t) \leq \text{FDP}_i^+(t)$ for all $t \in \mathbb{R}$ almost surely, which directly implies the ordering $T_i^- \leq T_i^* \leq T_i^+$.

Next, for each $i \in [m]$, we observe that

$$\begin{aligned} \frac{1}{K(n_{\text{cal}} + 1)} \sum_{j:(i,j) \in \mathcal{H}_0} \mathbb{E}[e_{i,j}] &= \mathbb{E} \left[\frac{\sum_{j:(i,j) \in \mathcal{H}_0} I(S_{n+i,j} \geq T_i^+)}{\sum_{k \in \mathcal{I}} \sum_{l=1}^K I(S_{k,l} \geq T_i^-) + \sum_{l=1}^K I(S_{n+i,l} \geq T_i^-) + \epsilon} \right] \\ &\leq \mathbb{E} \left[\frac{\sum_{j:(i,j) \in \mathcal{H}_0} I(S_{n+i,j} \geq T_i^*)}{\sum_{k \in \mathcal{I}} \sum_{l=1}^K I(S_{k,l} \geq T_i^*) + \sum_{l=1}^K I(S_{n+i,l} \geq T_i^*) + \epsilon} \right] \\ &\leq \mathbb{E} \left[\frac{\sum_{j=1}^K I(S_{n+i,j}^* \geq T_i^*)}{1 \vee \left\{ \sum_{k \in \mathcal{I}} \sum_{l=1}^K I(S_{k,l} \geq T_i^*) + \sum_{l=1}^K I(S_{n+i,l}^* \geq T_i^*) \right\}} \right] \leq \frac{1}{n_{\text{cal}} + 1}. \end{aligned}$$

Here, the first inequality arises from the ordering $T_i^- \leq T_i^* \leq T_i^+$, and the second inequality is directly implied by assumption (8). The final inequality follows from the hierarchical exchangeability of $\{S_k : k \in \mathcal{I}\} \cup \{S_{n+i}^*\}$. Therefore, by aggregating over all groups, we obtain

$$\sum_{(i,j) \in \mathcal{H}_0} \mathbb{E}[e_{i,j}] = \sum_{i=1}^m \sum_{j:(i,j) \in \mathcal{H}_0} \mathbb{E}[e_{i,j}] \leq mK,$$

which completes the proof. $\square$

B Other unit-level novelty detection methods

This appendix defines the full conformal and group-wise conformal extensions of HC-UND used in the numerical experiments. These procedures follow the same e-value construction as HSC-UND, but they fit the score function using more data. They are included as empirical extensions: unlike HSC-UND, their theoretical FDR control is not established in this paper because replacing non-null observations by hypothetical null observations can change fitted scores for reference groups and other test groups.

For HFC-UND, let $\mathcal{D} = \mathcal{D}_{\text{ref}} \cup \mathcal{D}_{\text{test}}$. We fit the location functional $C(z) = C(z; \mathcal{D})$ and the score function $s(x, x') = s(x, x'; \mathcal{D})$ using the full dataset, and define $S_{k,l} = s(X_{k,l}, C(Z_k))$ for $k \in [n + m]$ and $l \in [K]$. For each test group $i \in [m]$ and unit $j \in [K]$, the full conformal unit-level e-value is

$$e_{i,j}^{\text{FC}} = \frac{K(n+1)I(S_{n+i,j} \geq T_{i,\text{FC}}^+)}{\sum_{k \in [n] \cup \{n+i\}} \sum_{l=1}^K I(S_{k,l} \geq T_{i,\text{FC}}^-) + \epsilon}, \quad (10)$$

Algorithm 3 Hierarchical Full Conformal Unit-level Novelty Detection (HFC-UND)

Input: Reference data $\mathcal{D}_{\text{ref}} = \{X_{i,j} : i \in [n], j \in [K]\}$, test data $\mathcal{D}_{\text{test}} = \{X_{n+i,j} : i \in [m], j \in [K]\}$, target level α , threshold parameter $\tilde{\alpha}$.

- 1: Fit C and s using $\mathcal{D}_{\text{ref}} \cup \mathcal{D}_{\text{test}}$.
- 2: Compute $S_{k,l} = s(X_{k,l}, C(Z_k))$ for $k \in [n+m]$ and $l \in [K]$.
- 3: Compute $e_{i,j}^{\text{FC}}$ for $i \in [m]$ and $j \in [K]$ using (10).
- 4: Apply eBH to $\{e_{i,j}^{\text{FC}} : i \in [m], j \in [K]\}$.

Output: Selection set $\mathcal{R}_{\text{FC}} \subseteq [m] \times [K]$.

Algorithm 4 Hierarchical Group-wise Conformal Unit-level Novelty Detection (HGC-UND)

Input: Reference data $\mathcal{D}_{\text{ref}} = \{X_{i,j} : i \in [n], j \in [K]\}$, test data $\mathcal{D}_{\text{test}} = \{X_{n+i,j} : i \in [m], j \in [K]\}$, target level α , threshold parameter $\tilde{\alpha}$.

- 1: **for** $i = 1, \dots, m$ **do**
- 2: Fit $C^{(i)}$ and $s^{(i)}$ using $\mathcal{D}_i = \mathcal{D}_{\text{ref}} \cup \{X_{n+i,j} : j \in [K]\}$.
- 3: Compute $S_{k,l}^{(i)} = s^{(i)}(X_{k,l}, C^{(i)}(Z_k))$ for $k \in [n+m]$ and $l \in [K]$.
- 4: Compute $e_{i,j}^{\text{GC}}$ for $j \in [K]$ using (11).
- 5: **end for**
- 6: Apply eBH to $\{e_{i,j}^{\text{GC}} : i \in [m], j \in [K]\}$.

Output: Selection set $\mathcal{R}_{\text{GC}} \subseteq [m] \times [K]$.

where $T_{i,\text{FC}}^+ = \inf\{t \in [T_{\min}, \infty) : \text{FDP}_{i,\text{FC}}^+(t) \leq \tilde{\alpha}\}$ and $T_{i,\text{FC}}^- = \inf\{t \in [T_{\min}, \infty) : \text{FDP}_{i,\text{FC}}^-(t) \leq \tilde{\alpha}\}$, with

$$\begin{aligned} \text{FDP}_{i,\text{FC}}^+(t) &= \frac{m-1}{n+1} \frac{\sum_{k \in [n] \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j} \geq t) + \epsilon}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j} \geq t)}, \\ \text{FDP}_{i,\text{FC}}^-(t) &= \frac{m-1}{n+1} \frac{\sum_{k \in [n] \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j} \geq t)}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j} \geq t)}. \end{aligned}$$

For HGC-UND, the score function is fitted separately for each test group. For each $i \in [m]$, define $\mathcal{D}_i = \mathcal{D}_{\text{ref}} \cup \{X_{n+i,j} : j \in [K]\}$, fit $C^{(i)}(z) = C(z; \mathcal{D}_i)$ and $s^{(i)}(x, x') = s(x, x'; \mathcal{D}_i)$, and compute $S_{k,l}^{(i)} = s^{(i)}(X_{k,l}, C^{(i)}(Z_k))$ for $k \in [n+m]$ and $l \in [K]$. The group-wise conformal unit-level e-value is

$$e_{i,j}^{\text{GC}} = \frac{K(n+1)I(S_{n+i,j}^{(i)} \geq T_{i,\text{GC}}^+)}{\sum_{k \in [n] \cup \{n+i\}} \sum_{l=1}^K I(S_{k,l}^{(i)} \geq T_{i,\text{GC}}^-) + \epsilon}, \quad (11)$$

where $T_{i,\text{GC}}^+ = \inf\{t \in [T_{\min}, \infty) : \text{FDP}_{i,\text{GC}}^+(t) \leq \tilde{\alpha}\}$ and $T_{i,\text{GC}}^- = \inf\{t \in [T_{\min}, \infty) : \text{FDP}_{i,\text{GC}}^-(t) \leq \tilde{\alpha}\}$, with

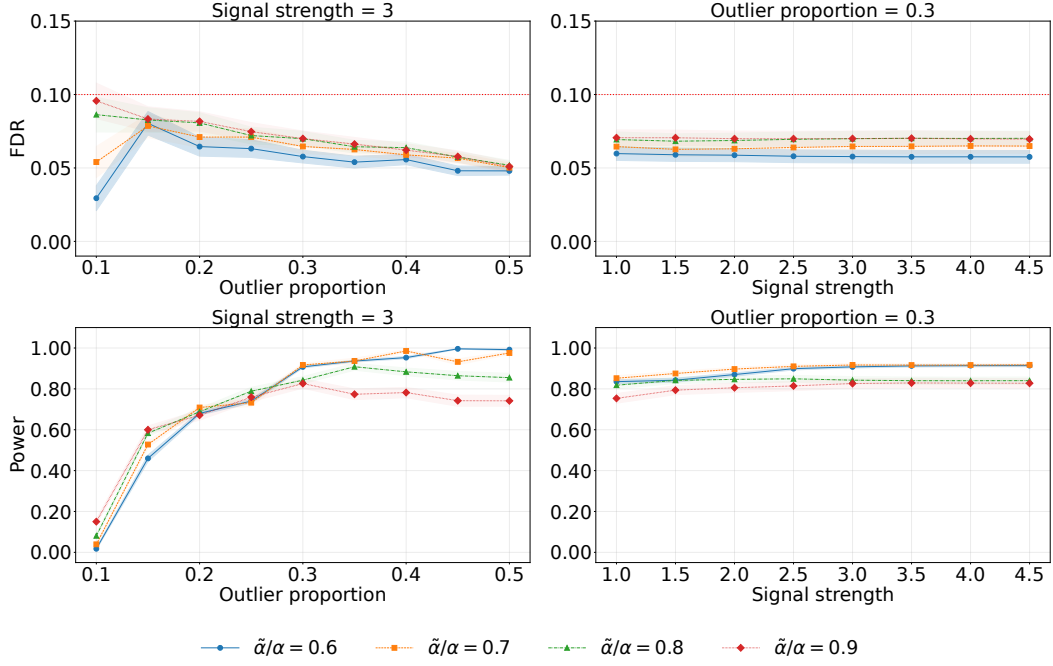


Figure 6: Sensitivity of HC-GND under the split conformal regime. The panels report empirical FDR and power as $\tilde{\alpha}/\alpha$ varies, with U-eBH used as the final multiple-testing step.

$$\text{FDP}_{i,\text{GC}}^+(t) = \frac{m-1}{n+1} \frac{\sum_{k \in [n] \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j}^{(i)} \geq t) + \epsilon}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j}^{(i)} \geq t)},$$

$$\text{FDP}_{i,\text{GC}}^-(t) = \frac{m-1}{n+1} \frac{\sum_{k \in [n] \cup \{n+i\}} \sum_{j=1}^K I(S_{k,j}^{(i)} \geq t)}{1 \vee \sum_{k \in [m] \setminus \{i\}} \sum_{j=1}^K I(S_{n+k,j}^{(i)} \geq t)}.$$

C Additional simulation results

C.1 Sensitivity analysis for selection of $\tilde{\alpha}/\alpha$

This appendix reports additional simulation results that support the sensitivity discussion in Section 3.5. We use the same simulation design as in Section 3.1 and vary $\tilde{\alpha}/\alpha$ while keeping the target level fixed at $\alpha = 0.1$. The outlier proportion is given by $\pi_1 \in \{0.1, 0.15, \dots, 0.5\}$, and the signal strength is given by $a \in \{1, 2, \dots, 4.5\}$ for both HC-GND and HC-UND. All experiments are repeated 1000 times, and the constant ϵ in the HC-UND e-value (6) is fixed at $\epsilon = 1$.

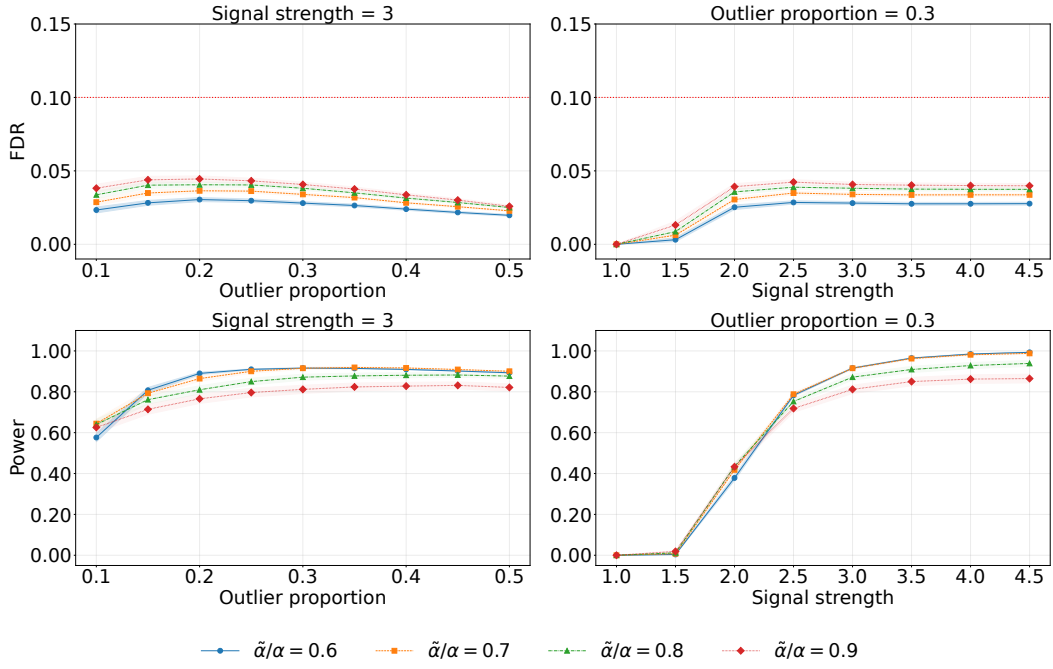


Figure 7: Sensitivity of HC-UND under the split conformal regime. The panels report empirical FDR and power as $\tilde{\alpha}/\alpha$ varies, with U-eBH used as the final multiple-testing step.

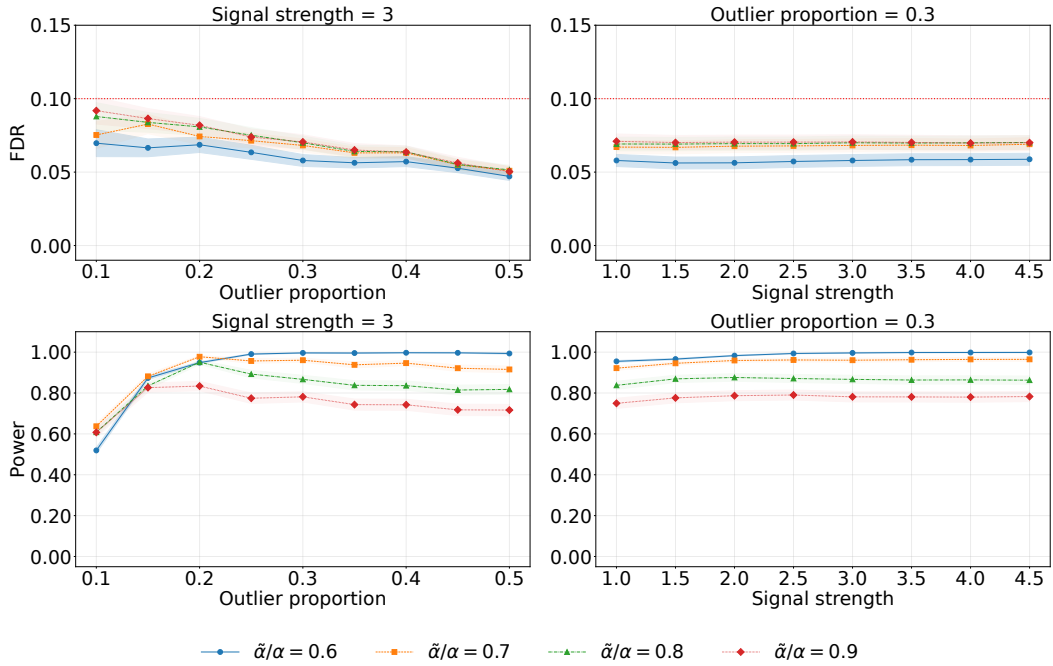


Figure 8: Sensitivity of HC-GND under the full conformal regime. The panels report empirical FDR and power as $\tilde{\alpha}/\alpha$ varies, with U-eBH used as the final multiple-testing step.

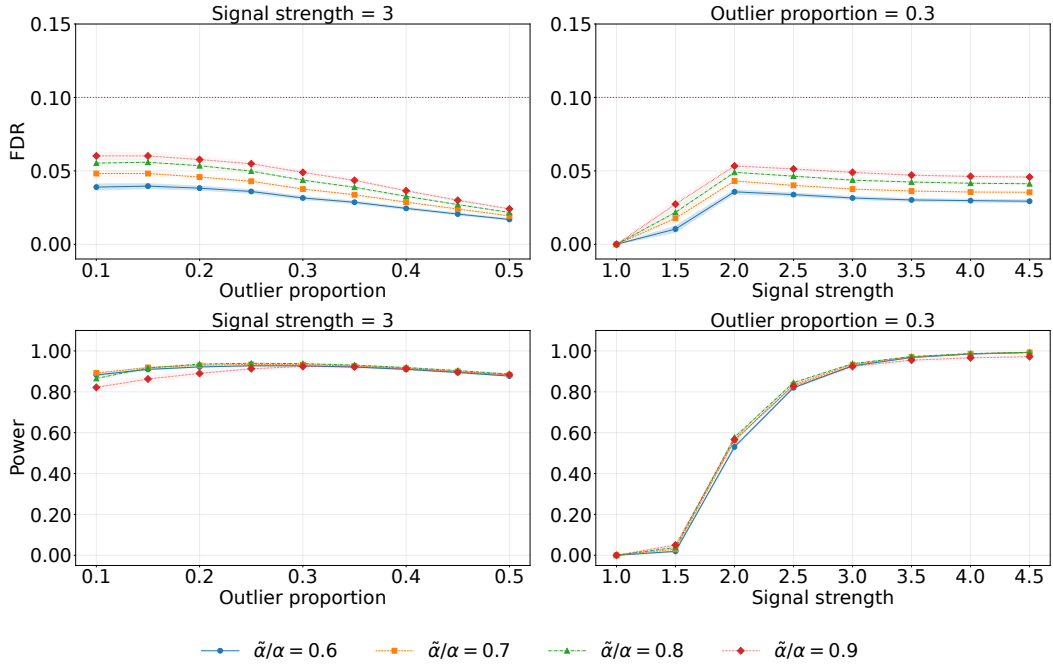


Figure 9: Sensitivity of HC-UND under the full conformal regime. The panels report empirical FDR and power as $\tilde{\alpha}/\alpha$ varies, with U-eBH used as the final multiple-testing step.

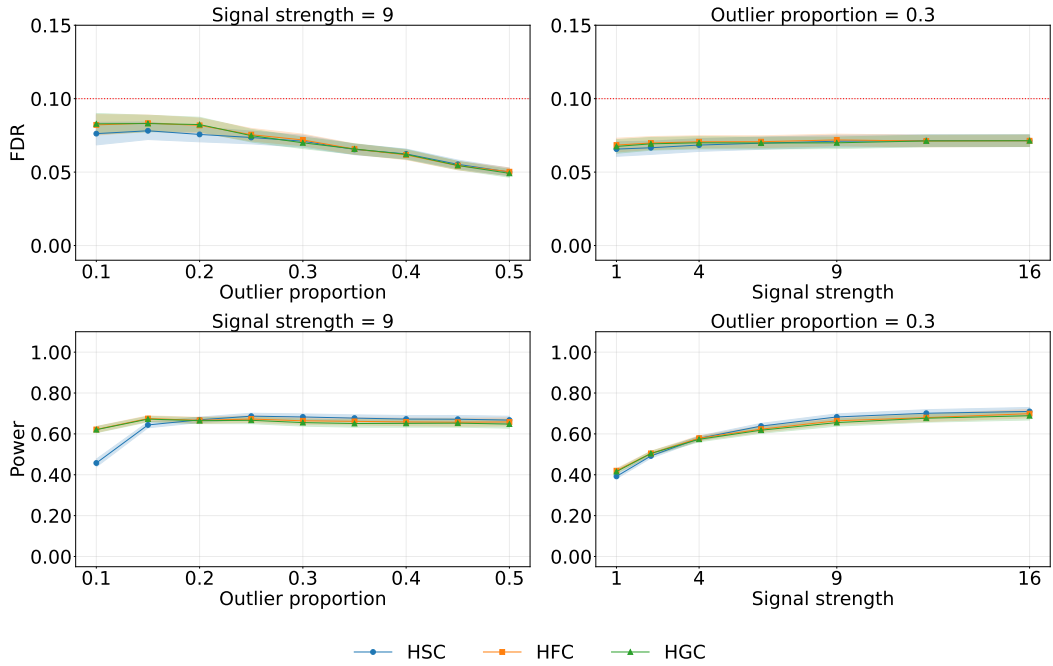


Figure 10: Additional low-dimensional simulation results for HC-GND. Empirical FDR and power are averaged over 1000 simulation runs under varying outlier proportions and signal strengths.

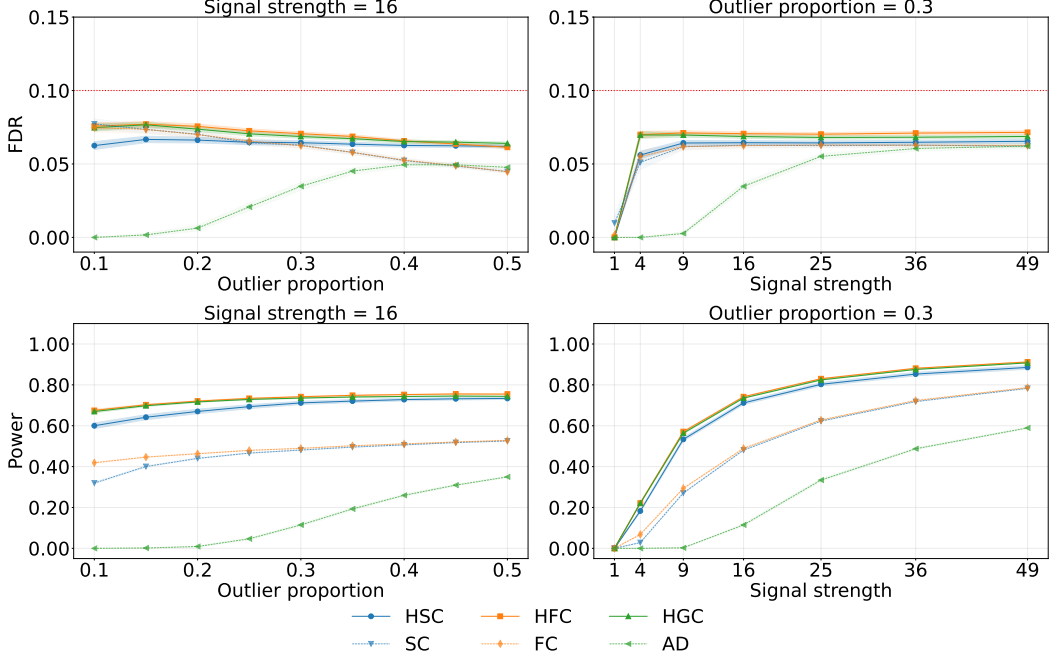


Figure 11: Additional low-dimensional simulation results for HC-UND and non-hierarchical baselines. Empirical FDR and power are averaged over 1000 simulation runs under varying outlier proportions and signal strengths.

C.2 Simulation setting reflecting real-world data

This subsection reports an additional simulation study motivated by the EV battery-pack analysis. The setting keeps the two-level hierarchy but uses a low-dimensional unit representation, reflecting the fact that the real data analysis applies novelty detection to two-dimensional FPCA score vectors. We compare the proposed HC-GND and HC-UND procedures with the non-hierarchical methods of Lee et al. (2026). The number of reference groups is fixed at $n = |\mathcal{D}_{\text{ref}}| = 400$, the number of test groups is fixed at $m = |\mathcal{D}_{\text{test}}| = 100$, each group contains $K = 10$ units, and each unit has dimension $d = 2$.

The signal strength is set to $a \in \{1^2, 1.5^2, \dots, 3.5^2, 4^2\}$ for HC-GND and $a \in \{1^2, 2^2, \dots, 7^2\}$ for HC-UND. Figure 10 reports the group-level results, and Figure 11 reports the unit-level comparison with the non-hierarchical baselines under the same outlier-generation model as in Section 3.4. This experiment is not intended to reproduce the EV dataset exactly; rather, it checks whether the qualitative simulation findings persist when the unit representation is low-dimensional, as in the FPCA-based data analysis.

The results are consistent with the main simulations. The empirical FDR of both HC-GND and HC-UND remains around or below the target level, while power increases as the signal becomes stronger. HC-GND maintains high empirical power across the

considered settings. For HC-UND, the proposed hierarchical methods retain a power advantage over the non-hierarchical SC, FC, and AdaDetect baselines. Compared with the higher-dimensional simulations in Figure 2, the power differences among HSC, HFC, and HGC are smaller, suggesting that the benefit of using additional data for score construction can be less pronounced when the unit-level representation is already low-dimensional.