A Unified Framework for Statistical Side-Channel Distinguishers: From Leakage Assessment to Key Recovery

Quentin L. Meunier

Sorbonne Université, LIP6, CNRS, 4 Place Jussieu 75005 Paris
quentin.meunier@lip6.fr

Abstract. Side-channel analysis (SCA) has historically developed along two parallel research avenues: key recovery attacks (e.g., CPA, MIA, Templates) and leakage assessment methodologies (e.g., TVLA, t-tests, F-tests, Kolmogorov-Smirnov tests). While both paradigms fundamentally process trace distributions partitioned by intermediate values or key hypotheses, they are typically viewed as disparate techniques with distinct mathematical formulations. In this paper, we propose a unifying statistical framework that decomposes side-channel distinguishers into orthogonal building blocks: (i) a subkey-induced leakage partition, (ii) a distribution comparison statistic (categorized into binary (two-distribution) or multi-class (multi-distribution) comparison operators), (iii) an aggregation operator over pairwise distribution comparisons, and (iv) a global decision rule. This formalization proves that classical key distinguishers and leakage tests are instances of a single generic model. Furthermore, our framework transforms statistical leakage tests into fully functional key recovery attacks, naturally introducing novel distinguishers based on non-parametric statistics (Kolmogorov-Smirnov, Anderson-Darling, MMD, Energy Distance) paired with suitable aggregation operators. We evaluate 18 statistical metrics across experimental configurations combining 8-bit AVR XMEGA and 32-bit Arm Cortex-M4 targets, unmasked and masked AES, under both targeted and automated points of interest scenarios. Our results demonstrate that the vast majority of these newly introduced attacks successfully achieve complete key recovery. Crucially, no single metric consistently dominates across all targets and leakage conditions, highlighting the complementary nature of statistical distinguishers and paving the way for metric combination and ensemble strategies.

Keywords: Side-channel analysis · Key distinguishers · Leakage assessment · Statistical hypothesis testing · Non-parametric statistics

1 Introduction

Side-Channel Analysis (SCA) allows adversaries to extract cryptographic keys from physical leakages such as power consumption and electromagnetic emissions [MOP07]. Over the past two decades, the SCA literature has evolved primarily along two distinct research avenues: *Side-Channel Attacks* and *Leakage Assessment Methodologies*. Side-Channel Attacks and Key Distinguishers are techniques designed to recover secret keys by identifying the candidate subkey hypothesis that maximizes the correlation or statistical dependency between physical measurements and intermediate cryptographic state predictions. Representative examples include Differential Power Analysis (DPA) [KJJ99], Correlation Power Analysis (CPA) [BCO04], Mutual Information Analysis (MIA) [GBSV08], and Template Attacks [CRR02]. Leakage Assessment Methodologies diagnostic procedures aimed at

detecting the presence of exploitable physical leakage without necessarily performing key recovery. These methods group trace measurements according to fixed versus random inputs or known key configurations, applying statistical goodness-of-fit or hypothesis testing tools such as Welch's t-test (TVLA) [GJJR11, SM15], Analysis of Variance (ANOVA) F-test [MS18], and Kolmogorov-Smirnov (KS) tests.

Despite their historical separation, both paradigms fundamentally manipulate empirical trace samples drawn from underlying probability distributions. In a key-recovery attack, a subkey hypothesis partitions the trace set into distribution classes corresponding to predicted intermediate states; the distinguisher then quantifies the separation or dependence across these classes. In leakage assessment, traces are similarly partitioned into two or more sets to evaluate distribution divergence.

Crucially, the literature frequently treats CPA, MIA, t-tests, F-tests, and distance metrics as a heterogeneous collection of independent distinguishers. There is a noticeable absence of a unified mathematical framework explaining how any statistical leakage test can be systematically converted into a robust key recovery distinguisher, and how different aggregation choices over multi-class partitions impact attack efficiency. We argue that the primary contribution of this work lies in a fundamental change of perspective: we demonstrate that all statistical side-channel key distinguishers can be factorized into four mutually orthogonal building blocks:

- **Leakage Partitioning**: A mapping induced by a subkey hypothesis $k \in \mathcal{K}$ that splits the trace universe into distribution classes $\mathbf{P}^{(k)} = \{P_0^{(k)}, \ldots, P_C^{(k)}\}$.
- **Statistical Comparison Function**: A core statistical operator S measuring separation or dependence. We establish a clear distinction between *binary (two-class) statistics* $S(P_i, P_j)$ and *multi-class statistics* $S(P_0, \ldots, P_C)$.
- **Aggregation Operator**: For binary pairwise statistics, a functional A combining the pairwise divergence metrics $\{S(P_i^{(k)}, P_j^{(k)})\}_{i<j}$ into a single scalar score (e.g., Arithmetic Mean or Maximum).
- **Global Decision Rule**: An optimization principle, typically $\hat{k} = \arg\max_{k \in \mathcal{K}} D(k)$, selecting the subkey hypothesis that maximizes the aggregated score.

This conceptual factorization provides a strong unifying thread. It explains why disparate metrics share theoretical bounds, reveals why certain aggregators fail under specific leakage models, and simplifies the creation of novel, highly effective key distinguishers from arbitrary statistical tests.

The main contributions of this paper are summarized as follows:

- We provide a unified statistical framework that unifies leakage assessment tests and classical key distinguishers into a single mathematical formulation. This framework proposes a classification separating binary statistics (e.g. t-statistic) from multi-class statistics (e.g. F-statistic).
- We define and analyze five aggregation operators for binary statistics, covering pairwise aggregations and two-class bipartition.
- We implemented the framework in a tool called `sca-metrics` which instantiates 18 metrics. This tool is provided online as open-source software[1].
- We evaluate these 18 metrics across experimental configurations combining 8-bit AVR XMEGA and 32-bit Arm Cortex-M4 microcontrollers, for unmasked and masked AES implementations.

[1] Code will be made available after acceptance of this work for publication

The rest of this paper is organized as follows: Section 2 reviews necessary background on side-channel analysis, cryptographic leakage models, and motivates the need for a unified perspective; Section 3 formalizes our unified statistical framework, detailing the four orthogonal building blocks and the taxonomy separating binary pairwise from multi-class operators; Section 4 presents the mathematical definitions and properties of the 18 instantiated metrics; Section 5 reports comprehensive experimental evaluations across 8-bit and 32-bit targets under both unmasked and masked AES implementations, analyzing aggregation strategies, sample efficiency, noise robustness, and computational execution costs; Finally, Section 6 discusses related work while Section 7 concludes the paper and outlines future directions.

2 Background and Motivation

In this section, we review classical side-channel key recovery attacks, leakage assessment techniques, and traditional definitions of side-channel distinguishers, laying the foundation for our unified framework.

2.1 Side-Channel Key Recovery Attacks

Consider an execution of a cryptographic algorithm processing a plaintext $p \in \mathcal{P}$ under a secret key $k^* \in \mathcal{K}$. During execution, the physical device produces a physical trace $T \in \mathbb{R}^D$ across D time samples. An adversary targets an intermediate state variable $V = f(p, k^*)$, where f is a known non-linear function (e.g., an AES S-box substitution $V = \text{SBox}[p \oplus k^*]$).

To recover k^*, the attacker selects a candidate key $k \in \mathcal{K}$ and predicts hypothetical intermediate variable $V^{(k)} = f(p, k)$ (with sample realizations $\{v_m^{(k)}\}_{m=1}^N$ for N traces). A side-channel distinguisher evaluates the statistical dependence between the set of N predicted values $\{v_m^{(k)}\}_{m=1}^N$ and the corresponding N observed physical leakage traces $\{T_m\}_{m=1}^N$. Classical attacks include Differential Power Analysis (DPA), Correlation Power Analysis (CPA) and Mutual Information Analysis (MIA). DPA [KJJ99] partitions traces according to a single or several bit(s) of $v^{(k)}$ into two sets and computes the difference of sample means; CPA [BCO04] assumes a linear power model (e.g., Hamming Weight $HW(v^{(k)})$) and estimates the Pearson correlation coefficient between predicted weights and trace amplitudes; while MIA [GBSV08] uses probability density estimation (histograms or kernels) to measure the mutual information $I(T; V^{(k)})$ without assuming a specific power model.

2.2 Leakage Assessment Methodologies

Rather than recovering secret subkeys, leakage assessment methodologies evaluate whether physical side-channel traces contain statistically significant information related to sensitive operations. Classical tests include the Welch's t-test, the ANOVA F-test, and the Kolmogorov-Smirnov (KS) Test. The Welch's t-test [GJJR11, SM15] or Test Vector Leakage Assessment (TVLA) methodology compares two sets of traces—typically fixed vs. random inputs—using Welch's t-statistic to test the null hypothesis $H_0 : \mu_0 = \mu_1$. A threshold of $|t| > 4.5$ indicates physical leakage for a given sample with 99.999% confidence. The ANOVA F-test [MS18] generalizes TVLA to multi-class partitions by comparing the variance between class means to the average variance within classes. Finally, the KS Test evaluates the maximum distance between cumulative distribution functions (CDFs) of two trace sets, providing a non-parametric alternative to Welch's t-test. Crucially, every leakage assessment tool is fundamentally a *statistical test* applied to empirical trace distributions.

2.3 Classical Distinguishers vs. Generalized View

In classical literature, a side-channel distinguisher is defined as a mapping:

$$\mathcal{D} : \mathcal{K} \times (\mathbb{R}^D \times \mathcal{P})^N \to \mathbb{R}, \tag{1}$$

assigning a numerical score to each candidate key $k \in \mathcal{K}$. The candidate maximizing $\mathcal{D}(k)$ is chosen as the estimated key $\hat{k}$.

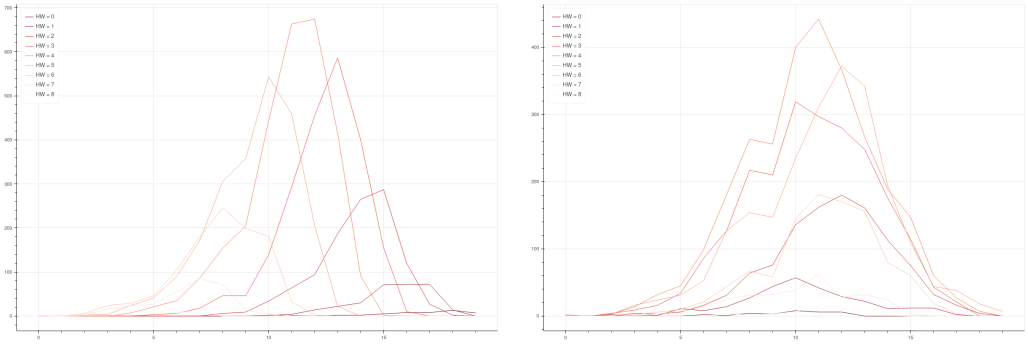
While functional, this classical definition treats $\mathcal{D}(k)$ as a black box. It masks the internal mechanics: how traces are grouped, how distributions are compared, and how multiple pairwise comparisons are combined when the intermediate state takes $C + 1 > 2$ values. In the next section, we open this black box to establish a comprehensive, modular framework.

3 Unified Statistical Framework

This section presents the main theoretical contribution of our paper: a generic, modular framework that formalizes side-channel distinguishers into orthogonal constituents.

Intuition and Motivation. The core principle of partition-based side-channel analysis relies on a fundamental observation: partitioning physical trace measurements based on a subkey hypothesis $k \in \mathcal{K}$ yields distinct probability distributions if and only if k matches the correct secret key k^*.

When the correct key hypothesis is chosen ($k = k^*$), the partition aligns with the physical states of the device. Consequently, the traces are grouped into distinct distributions $\{P_0, \ldots, P_C\}$ corresponding to each leakage class (e.g., the 9 Hamming Weight classes), which exhibit visible separation and distinct characteristics, as illustrated in Figure 1(a). Conversely, when a wrong key hypothesis is chosen ($k \neq k^*$), the partition maps plaintexts to leakage classes randomly. The resulting distributions are mixtures of all actual physical states, causing them to superimpose and overlap completely, as shown in Figure 1(b).



(a) Correct key hypothesis $k = k^*$ (b) Wrong key hypothesis $k \neq k^*$

Figure 1: Empirical distributions of power consumption at the POI for an intermediate value, partitioned by Hamming Weight class under the correct and wrong key hypotheses.

The unified framework proposed in this section formalizes this intuition: by quantifying the statistical discrepancy across these distributions, it shows how to systematically transform any leakage assessment test, statistical distance, or divergence metric into a full-fledged key-recovery distinguisher.

3.1 Leakage Partitioning

Let $\mathcal{T} = \{(T_m, p_m)\}_{m=1}^{N}$ denote a set of N traces paired with plaintexts. Given a target intermediate function $f(p, k)$ and a leakage model ϕ, each key hypothesis $k \in \mathcal{K}$ induces a partition of the trace set $\mathcal{T}$ into $C + 1$ classes:

$$\mathcal{T}_c^{(k)} = \{T_m \in \mathcal{T} \mid \phi(f(p_m, k)) = c\}, \quad c \in \{0, \ldots, C\}. \tag{2}$$

Let $P_c^{(k)}$ denote the empirical probability distribution of traces in class $\mathcal{T}_c^{(k)}$, and $n_c^{(k)} = |\mathcal{T}_c^{(k)}|$ denote the class sample count.

For example, when targeting an 8-bit intermediate state $V \in \mathbb{F}_2^8$ under the Hamming Weight model $\phi(V) = HW(V)$, the universe is partitioned into $C + 1 = 9$ classes $\{P_0^{(k)}, \ldots, P_8^{(k)}\}$; while when targeting the full byte value $\phi(V) = V$, the universe is partitioned into $C + 1 = 256$ classes $\{P_0^{(k)}, \ldots, P_{255}^{(k)}\}$.

3.2 Statistical Comparison Functions

Once traces are partitioned into distributions $\mathbf{P}^{(k)} = \{P_0^{(k)}, \ldots, P_C^{(k)}\}$, a statistical function S evaluates their divergence or dependency. We divide statistical functions into two fundamental categories:

Definition 1 (Binary Statistical Function)**.** A binary statistical function $S_{\text{bin}} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}$ evaluates a pairwise distance, divergence, or test statistic between two probability distributions P_i and P_j:

$$S_{ij} = S_{\text{bin}}(P_i, P_j). \tag{3}$$

Examples include Welch's t-statistic, Kolmogorov-Smirnov distance, Anderson-Darling statistic, Energy Distance, Maximum Mean Discrepancy, and Difference of Means.

Definition 2 (Multi-Class Statistical Function)**.** A multi-class statistical function $S_{\text{multi}} : \mathcal{P}^{C+1} \to \mathbb{R}_{\geq 0}$ simultaneously evaluates the joint divergence or global variance across all $C + 1$ distributions:

$$S_{\text{multi}}(\mathbf{P}^{(k)}) = S_{\text{multi}}(P_0^{(k)}, \ldots, P_C^{(k)}). \tag{4}$$

Examples include the ANOVA F-statistic, Mutual Information Analysis, Kruskal-Wallis statistic, and Variance of Means.

In addition, classical correlation metrics (e.g., Pearson, Spearman) can be viewed as dependence functions directly acting on the bivariate sample distribution $(Y, X^{(k)})$ where $X^{(k)} = \phi(f(p, k))$.

3.3 Aggregation Operators for Binary Statistics

When a binary statistic S_{bin} is applied to a multi-class partition ($C + 1 > 2$), it yields a symmetric matrix of $N_{\text{pairs}} = \binom{C+1}{2} = \frac{C(C+1)}{2}$ pairwise values:

$$\mathbf{S}^{(k)} = \left\{ S_{ij}^{(k)} = S_{\text{bin}}(P_i^{(k)}, P_j^{(k)}) \;\middle|\; 0 \leq i < j \leq C \right\}. \tag{5}$$

To collapse this matrix into a single scalar score for key guess k, we define an *aggregation operator* to be a symmetric and monotonic operator $A : \mathbb{R}_{\geq 0}^{N_{\text{pairs}}} \to \mathbb{R}_{\geq 0}$. We choose to study five aggregation operators (four pairwise operators and a two-class bipartition operator), each translating a specific physical or statistical assumption regarding the underlying leakage:

- **Arithmetic Mean (`MEAN`)** assumes a homogeneous leakage profile across all state transitions. By averaging over all N_{pairs} combinations, it maximizes robustness against additive Gaussian noise by leveraging the law of large numbers.

$$A_{\text{mean}}(\mathbf{S}^{(k)}) = \frac{1}{N_{\text{pairs}}} \sum_{0 \le i < j \le C} S_{ij}^{(k)}. \tag{6}$$

- **Weighted Mean (`WMEAN`)** accounts for class population imbalances (e.g., the binomial distribution of the Hamming Weight model). By weighting each pair by $w_{ij}^{(k)} = \frac{n_i^{(k)} n_j^{(k)}}{n_i^{(k)} + n_j^{(k)}}$, it minimizes the statistical influence of high-variance, low-sample pairs (such as $HW \in \{0, 8\}$ comparisons), speeding up convergence.

$$A_{\text{wmean}}(\mathbf{S}^{(k)}) = \frac{\sum_{0 \le i < j \le C} w_{ij}^{(k)} S_{ij}^{(k)}}{\sum_{0 \le i < j \le C} w_{ij}^{(k)}}. \tag{7}$$

- **Root Mean Square (`RMS`)** is physically motivated by energy-based leakages. It penalizes smaller differences while highlighting stronger pairwise discrepancies, acting as a smoother alternative to the maximum operator.

$$A_{\text{rms}}(\mathbf{S}^{(k)}) = \sqrt{\frac{1}{N_{\text{pairs}}} \sum_{0 \le i < j \le C} \left(S_{ij}^{(k)}\right)^2}. \tag{8}$$

- **Maximum (`MAX`)** represents a worst-case scenario. It is mathematically optimal when the leakage is sparse and highly localized to a single transition pair, preventing the informative signal from being diluted by averaging over non-leaking pairs. However, it is highly sensitive to extreme noise outliers.

$$A_{\max}(\mathbf{S}^{(k)}) = \max_{0 \le i < j \le C} S_{ij}^{(k)}. \tag{9}$$

- **Two-Class Bipartition (`TCB`)** generalizes the classical multi-bit DPA principle to any binary metric. Instead of computing $\binom{C+1}{2}$ pairwise comparisons across sparse individual classes, it partitions the trace set upstream into two well-populated macro-classes around the median value:

$$P_<^{(k)} = \bigcup_{c < \text{mid}} P_c^{(k)}, \quad P_>^{(k)} = \bigcup_{c > \text{mid}} P_c^{(k)}, \tag{10}$$

discarding the uninformative central partition ($c = \text{mid}$). The distinguisher score is directly obtained by evaluating the binary statistic between these two macro-distributions:

$$A_{\text{tcb}}(\mathbf{S}^{(k)}) = S_{\text{bin}}\left(P_<^{(k)}, P_>^{(k)}\right). \tag{11}$$

This operator maximizes the sample size per evaluated distribution while maximizing the theoretical contrast between the lower and upper halves of the intermediate state.

3.4 Final Distinguisher

Combining the components yields a single general formula for the distinguisher score $D(k)$:

$$D(k) = \begin{cases} A\left(\left\{S_{\text{bin}}(P_i^{(k)}, P_j^{(k)})\right\}_{i<j}\right) & \text{if } S \text{ is binary,} \\ S_{\text{multi}}\left(P_0^{(k)}, \ldots, P_C^{(k)}\right) & \text{if } S \text{ is multi-class.} \end{cases} \tag{12}$$

The subkey guess $\hat{k}$ is extracted via the standard optimization rule:

$$\hat{k} = \arg\max_{k \in \mathcal{K}} D(k). \tag{13}$$

4 Metric Instantiations

In this section, we present the mathematical definitions, statistical contexts, and typical side-channel use cases of the 18 statistical functions evaluated within our unified framework.

4.1 Dependence Metrics

Pearson Correlation Coefficient (`P-corr`). It evaluates the strength of a linear dependency between trace values Y and predictions $X^{(k)}$:

$$\rho(Y, X^{(k)}) = \frac{\sum(Y - \bar{Y})(X^{(k)} - \bar{X}^{(k)})}{\sqrt{\sum(Y - \bar{Y})^2 \sum(X^{(k)} - \bar{X}^{(k)})^2}} \tag{14}$$

Use Case: This is the standard distinguisher in Correlation Power Analysis (CPA) [BCO04]. It is theoretically optimal when the physical leakage is linear with respect to the intermediate model (e.g., Hamming Weight on microcontrollers). Its primary advantage is computational efficiency, as it can be computed incrementally over millions of traces.

Spearman Rank Correlation (`S-corr`). It corresponds to the Pearson correlation computed on the ranks $R(Y)$ and $R(X^{(k)})$:

$$\rho_S(Y, X^{(k)}) = \rho(R(Y), R(X^{(k)})) \tag{15}$$

Use Case: It is typically used when the physical leakage function is monotonic but non-linear, or when the measurement noise is heavy-tailed with outliers (sorting ranks dampens the effect of extreme outliers). It is highly suitable for black-box evaluations where the precise physical leakage profile is unknown.

4.2 Multi-Class Statistical Functions

ANOVA F-Statistic (`F-stat`). It evaluates the ratio of between-class variance to within-class variance across the $C + 1$ partitioned groups:

$$F = \frac{\frac{1}{C}\sum_{c=0}^{C} n_c(\hat{\mu}_c - \hat{\mu})^2}{\frac{1}{N-C-1}\sum_{c=0}^{C}\sum_{x \in \mathcal{T}_c}(x - \hat{\mu}_c)^2} \tag{16}$$

Use Case: In standard statistics, the F-test evaluates the null hypothesis H_0 that all $C + 1$ classes originate from distributions with identical means under the assumption of normality and homoscedasticity. In SCA, it is standard for leakage assessment (ANOVA test). When instantiated as a key distinguisher, it acts as a model-independent tool that identifies if a subkey hypothesis creates classes with differing means, without assuming a linear relation (unlike CPA).

Model-Based Mutual Information Analysis (`MIA-mb`). It quantifies the mutual dependency $I(Y; X^{(k)}) = H(Y) - H(Y|X^{(k)})$ under a parametric Gaussian assumption for the leakage distribution, where the conditional entropy of each class is weighted according to the theoretical binomial probabilities of the Hamming Weight distribution.

Use Case: It is typically used when the physical leakage follows a known theoretical probability model, stabilizing the estimation of entropy for rarely-occurring classes in low-trace count environments.

Query-Based Mutual Information Analysis (`MIA-qb`). Similarly to `MIA-mb`, it estimates the conditional entropy under a parametric Gaussian assumption, but weights each class by its empirical query count within the active trace subset instead of theoretical model probabilities.

Use Case: A model-independent version of MIA that adapts to the empirical distribution of the captured traces, making it highly robust to imbalances in plaintext distributions.

Kruskal-Wallis Statistic (`KW-stat`). A non-parametric multi-class rank sum test evaluating whether the medians of the $C+1$ classes are equal:

$$H = \frac{12}{N(N+1)} \sum_{c=0}^{C} \frac{R_c^2}{n_c} - 3(N+1) \tag{17}$$

where R_c is the sum of ranks in group c.
Use Case: Serving as a non-parametric alternative to ANOVA, the Kruskal-Wallis test is used when trace distributions deviate significantly from normality, or when the measurement setup introduces non-Gaussian noise and outliers that would bias the F-statistic.
Variance of Means (`VoM`). It computes the sample variance of the class means:

$$VoM = \frac{1}{C} \sum_{c=0}^{C} (\hat{\mu}_c - \bar{\mu})^2 \tag{18}$$

Use Case: Standard metric in vertical SCA representing the physical signal amplitude. Under a linear leakage model, `VoM` is directly proportional to the square of the Pearson correlation coefficient.
Min Mean of Variances (`MMoV`) and **Min of Max Variances (`MoMV`)**. These metrics evaluate the variances within each partition class:

$$MMoV = \frac{1}{C+1} \sum_{c=0}^{C} \hat{\sigma}_c^2, \quad MoMV = \max_{0 \le c \le C} \hat{\sigma}_c^2 \tag{19}$$

Use Case: These metrics target second-order leakage. For the correct key guess, partitioning traces according to the correct intermediate value yields classes with minimum internal variance (as the signal component is constant within each class). For incorrect key guesses, the classes are random mixtures, increasing the internal variance.

4.3 Parametric Binary Distance Metrics

Difference of Means (`DoM`) and **Welch's t-Statistic (`T-stat`)**. They measure respectively the unstandardized and standardized difference of means between two distributions P_i and P_j:

$$DoM_{ij} = |\hat{\mu}_i - \hat{\mu}_j|, \quad t_{ij} = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{\sqrt{\frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_j^2}{n_j}}} \tag{20}$$

Use Case: `DoM` is the classic metric for DPA. `T-stat` is the standard metric used in TVLA to detect leakage under unequal class variances. Within our framework, aggregating these pairwise metrics transforms these leakage detection tests into full key recovery distinguishers.

4.4 Non-Parametric Binary Distance Metrics

Kolmogorov-Smirnov Statistic (`KS-stat`). It evaluates the supremum distance between empirical cumulative distribution functions (CDFs) $F_i(x)$ and $F_j(x)$:

$$D_{KS}(P_i, P_j) = \sup_x |F_i(x) - F_j(x)| \tag{21}$$

Use Case: The KS-test is non-parametric and sensitive to any difference in distribution shape (location, scale, or shape). It is used in SCA to detect higher-order leakages in masked implementations where first-order means are equal.

Anderson-Darling Statistic (`AD-stat`). A modification of the KS-statistic that places higher weight on the tails of the distributions. Evaluated across the pooled and sorted sample sequence $\{z_k\}_{k=1}^{N_{ij}-1}$ (with $N_{ij} = n_i + n_j$), the empirical two-sample statistic is computed as:

$$A_{ij}^2 = \frac{n_i n_j}{N_{ij}} \sum_{k=1}^{N_{ij}-1} \frac{\left(F_i(z_k) - F_j(z_k)\right)^2}{\frac{k}{N_{ij}}\left(1 - \frac{k}{N_{ij}}\right)} \tag{22}$$

where F_i and F_j denote the empirical CDFs of both classes.
Use Case: It is highly sensitive to differences in the tails of the distributions, which are common in high-frequency side-channel measurements where rare, large-amplitude power transitions carry critical leakage.
Maximum Mean Discrepancy (`MMD-stat`). A kernel-based distance [GBR+12] measured in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with a Gaussian kernel $k(x, x') = \exp(-\gamma(x - x')^2)$. Given sample sets $\{x_a\}_{a=1}^{n_i} \sim P_i$ and $\{y_b\}_{b=1}^{n_j} \sim P_j$, the empirical statistic is evaluated as:

$$\mathrm{MMD}^2(P_i, P_j) = \frac{1}{n_i^2} \sum_{a,a'=1}^{n_i} k(x_a, x_{a'}) + \frac{1}{n_j^2} \sum_{b,b'=1}^{n_j} k(y_b, y_{b'}) - \frac{2}{n_i n_j} \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} k(x_a, y_b) \tag{23}$$

Use Case: `MMD-stat` is capable of capturing infinite-dimensional moment differences. It is used as a powerful non-parametric distinguisher that operates without prior density estimation, and is particularly suited for multivariate leakage profiles.
Energy Distance (`E-dist`). A statistical distance [SR04] based on pairwise Euclidean distances between sample points:

$$D_E^2(P_i, P_j) = \frac{2}{n_i n_j} \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} |x_a - y_b| - \frac{1}{n_i^2} \sum_{a,a'=1}^{n_i} |x_a - x_{a'}| - \frac{1}{n_j^2} \sum_{b,b'=1}^{n_j} |y_b - y_{b'}| \tag{24}$$

Use Case: Similar to `MMD-stat`, it is a robust non-parametric distance between distributions. It avoids the need to tune kernel parameters (such as the Gaussian bandwidth in MMD) and behaves as a scale-invariant metric.

4.5 Information Divergence Metrics

Bhattacharyya Distance (`B-dist`). It measures the overlap between empirical probability distributions partitioned into B histogram bins (where $p_i(b)$ is the relative frequency in bin b):

$$D_B(P_i, P_j) = -\ln \sum_{b=1}^{B} \sqrt{p_i(b)\, p_j(b)} \tag{25}$$

Use Case: Used to quantify the separability of trace distributions when the distributions are non-Gaussian but share a similar shape.
Kullback-Leibler (`KL-div`) and **Jensen-Shannon Divergences (`JS-div`)**. They evaluate the asymmetric relative entropy and symmetric divergence across the B discrete bins:

$$D_{KL}(P_i \parallel P_j) = \sum_{b=1}^{B} p_i(b) \ln \frac{p_i(b)}{p_j(b)} \tag{26}$$

$$D_{JS}(P_i, P_j) = \frac{1}{2} D_{KL}(P_i \parallel M) + \frac{1}{2} D_{KL}(P_j \parallel M) \tag{27}$$

where $M = \frac{1}{2}(P_i + P_j)$.
Use Case: These metrics quantify the information leakage between class distributions. The JS-divergence is symmetric and bounded in $[0, \ln(2)]$, making it statistically stable for pairwise aggregation.

5 Experimental Results

5.1 Methodology

To validate our unified framework and compare the efficiency of the different statistical functions and aggregation operators, we implemented the designed metrics in a tool called `sca-metrics` and computed them for different configurations. Specifically, we targeted two hardware target platforms, which contain representative embedded microcontrollers commonly used in side-channel benchmark suites:

- **Atmel AVR XMEGA128D4**: An 8-bit RISC architecture characterized by clear, high signal-to-noise ratio (SNR) power consumption leakages.

- **Arm Cortex-M4 (32-bit)**: A modern 32-bit Arm architecture featuring pipeline stages, bus multiplexing, and higher algorithmic noise.

For each hardware platform, we targeted two AES-128 software implementations:

- `AES-128`: Standard, unprotected AES software execution, using a tabulated SBox. The target intermediate value is the first-round S-Box output $V = \text{SBox}[p_0 \oplus k_0^*]$.

- `Masked AES-128`: First-order Boolean masked AES implementation using the scheme defined in [HOM06] where intermediate variables are protected by a random secret mask M: $V_{\text{masked}} = \text{SBox}[p_0 \oplus k_0^*] \oplus M$. Hence, each intermediate variable has a uniform distribution.

For each configuration, we recorded power traces with the ChipWhispererPro device [OC14] and target boards AVR XMEGA and STM32F303, using 4 samples per cycles. The capture power traces correspond to the first round of the AES, with variable plaintexts and a fixed key. The traces are available online [2].

We also considered two distinct scenarios regarding the Points of Interest (POIs) in order to assess the robustness of metrics against noise outside target clock cycles:

- **POI-Selected scenario**: Statistical metrics are evaluated strictly at predefined POIs corresponding to target S-Box clock cycles. Those are selected to maximize the correlation between power and the Hamming Weight of the target intermediate value.

- **POI-Auto scenario**: Metrics are computed across the entire trace timeline without prior knowledge of POI locations, and the chosen sample for each metric and byte is the one maximizing the metric for the total number of traces.

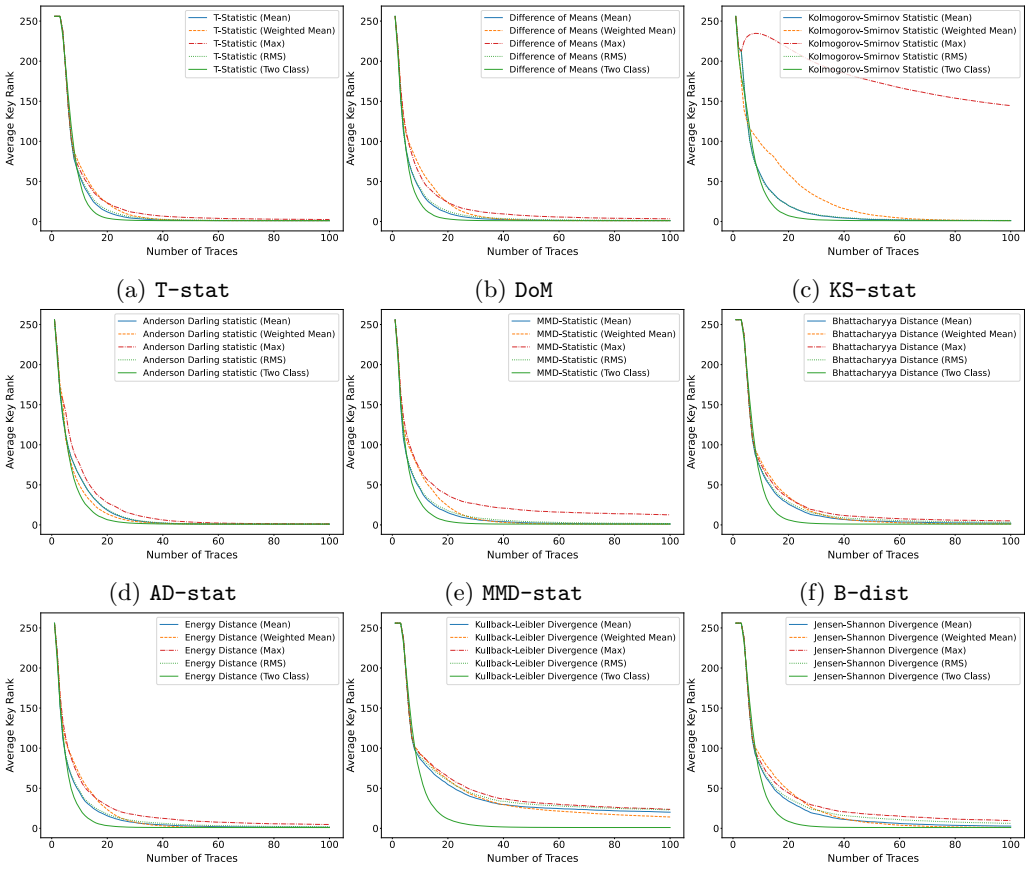
For each setting, we compute the *Average Key Rank* of the secret key hypothesis k^* as a function of the trace count N. The key rank KR(N) $\in [1, 256]$ measures the number of candidate subkeys scoring strictly lower than k^*. A KR of 1 means that k^* has the best score and that the attack is successful. An average KR of 1 means that all 16 bytes are successfully recovered. Moreover, in order to ensure statistical significance, an `averaging` parameter has been used to perform the attack on independent trace subsets, whose results have then been averaged together for each number of traces. For example, with `averaging` = 100, an average KR of 1 means that the correct key is found for all bytes in all 100 experiments.

The metrics computation have been run on the Dalek cluster [CAB25].

[2] They will be made available after acceptance of this work for publication

5.2 Comparison of Aggregation Operators

We first evaluate the five aggregation operators (`MEAN`, `WMEAN`, `RMS`, `MAX`, and `TCB`) applied to binary statistics. For this, we use the POI-Selected scenario with `AES-128` on the AVR XMEGA target, with an `averaging` value of 100, and evaluate the average key rank as a function of the number of traces. Results are shown in Figure 2, while the metric values for 100 traces are shown in Table 1.



(a) `T-stat` (b) `DoM` (c) `KS-stat`

(d) `AD-stat` (e) `MMD-stat` (f) `B-dist`

(g) `E-dist` (h) `KL-div` (i) `JS-div`

Figure 2: Comparison of Key Rank convergence across the nine evaluated binary statistical metrics, for the AVR XMEGA target and with an `averaging` value of 100 and the POI-Selected scenario.

As shown in Figure 2 and Table 1, the Two-Class Bipartition (`TCB`) operator overwhelmingly outperforms all pairwise aggregation operators across all nine binary metrics. At $N = 100$ traces, `TCB` drives the average key rank to the optimal value of 1.0000 for eight metrics, and to 1.0006 for `KL-div` (which remained above 14.18 with pairwise `WMEAN`). Even more importantly, `TCB` delivers drastically faster convergence in the critical early-trace regime ($N \leq 30$). For example, with `DoM` at $N = 15$ traces, `TCB` already reaches an average key rank of 7.39, compared to 19.22 for `MEAN` and 41.76 for `WMEAN`.

This marked superiority stems from two statistical factors inherent to macro-class pooling: first, pairwise aggregation divides N traces into $C + 1 = 9$ fine-grained partitions. In the low-trace regime, extreme classes ($HW \in \{0, 1, 7, 8\}$) contain virtually zero or only a couple of traces, introducing heavy sampling variance across the $\binom{9}{2} = 36$ pairwise estimators. In contrast, `TCB` pools traces into two large, balanced macro-classes ($HW < 4$

Table 1: Average Key Rank for 100 traces on AVR XMEGA target for binary statistics across aggregation operators with an `averaging` value of 100.

	MEAN	WMEAN	RMS	MAX	TCB	**Selected**
DoM	1.0619	**1.0000**	1.3288	3.3694	**1.0000**	TCB
T-stat	1.0219	**1.0000**	1.1987	2.5125	**1.0000**	TCB
KS-stat	1.1425	1.0925	1.2956	144.5469	**1.0000**	TCB
AD-stat	1.0012	**1.0000**	1.0025	1.1694	**1.0000**	TCB
E-dist	1.3081	1.0025	2.0000	4.8006	**1.0000**	TCB
MMD-stat	1.2169	1.0025	1.7406	12.6375	**1.0000**	TCB
B-dist	2.0425	1.1706	3.3281	5.0381	**1.0000**	TCB
KL-div	20.1187	14.1831	23.0381	23.6400	**1.0006**	TCB
JS-div	2.7462	1.1612	6.1750	9.7088	**1.0000**	TCB

and $HW > 4$), each concentrating $\approx 36.3\%$ of all available traces, thereby drastically reducing empirical estimation error. This benefit is particularly critical for non-parametric and distribution-based metrics (such as `KL-div`, `JS-div`, and `B-dist`), whose density and histogram estimators require sufficient sample counts per class to yield reliable statistics.

Second, by discarding the median partition ($HW = 4$), which comprises $\approx 27.3\%$ of the traces that exhibit minimal contrast relative to the mean, `TCB` eliminates uninformative noise and maximizes the theoretical mean contrast between the two evaluated distributions ($\Delta\overline{HW} \approx 3.02$ bits).

Among the classical pairwise aggregation operators, our results reveal a distinct crossover effect between `MEAN` and `WMEAN` in the low-trace regime ($N \leq 30$). For instance, for `DoM` at $N = 10$ traces, the unweighted `MEAN` achieves a key rank of 39.91, whereas `WMEAN` lags at 67.15 (compared to 23.43 for `TCB`), with `WMEAN` only overtaking `MEAN` after approximately 50 traces (1.50 vs. 1.66). This behavior is governed by the binomial distribution of the Hamming Weight model: the strongest contrast occurs between extreme classes ($\Delta HW = 8$), which are rare ($p = 1/256$). When N is small, their sample counts are negligible, causing `WMEAN` (which weights pairs by $w_{ij} = \frac{n_i n_j}{n_i + n_j}$) to heavily penalize these high-contrast pairs and concentrate its weight on central, low-contrast pairs ($\Delta HW = 1$). In contrast, `MEAN` preserves the contribution of high-contrast pairs regardless of sample size, improving early convergence. Once $N \geq 50$ and all classes are adequately populated, `WMEAN` minimizes overall estimation variance.

The Root Mean Square (`RMS`) operator behaves similarly to `MEAN` (reaching a key rank of 41.78 at $N = 10$ for `DoM`). By squaring pairwise distances, `RMS` emphasizes strong discrepancies, but this quadratic weighting also magnifies the estimation variance of sparse classes when N is small. Finally, the Maximum (`MAX`) operator consistently exhibits the worst performance across all evaluations (e.g., reaching only 3.37 for `DoM` and failing entirely for `KS-stat` at 144.55 at $N = 100$). Because `MAX` selects the extreme value over 36 comparisons, under incorrect key hypotheses it captures the largest positive noise spike, entirely forfeiting the noise-averaging benefits of the law of large numbers.

Given that `TCB` delivers superior convergence across all evaluated binary metrics in both early and asymptotic regimes, it is retained as the optimal aggregator for binary statistics in the subsequent experiments.

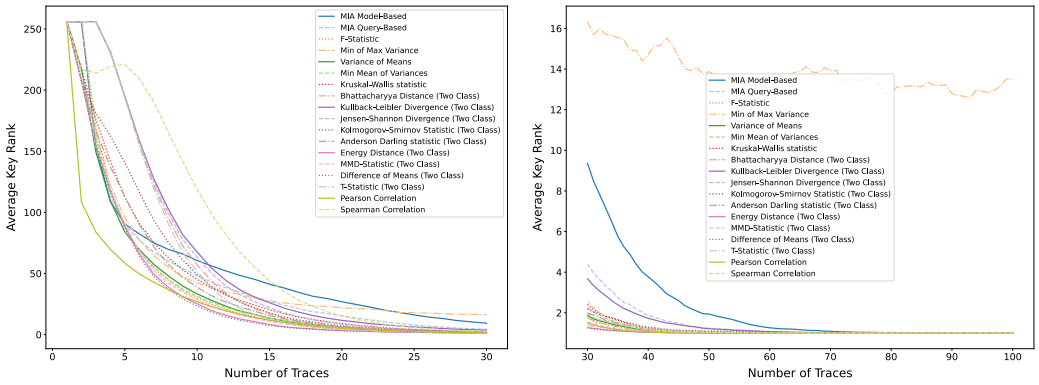
5.3 Key Recovery Performance on Unprotected AES-128

In this section, we analyze the sample efficiency and key rank convergence of the 18 statistical metrics on the unprotected software AES-128 implementation, with the POI-Selected scenario. We target both the 8-bit AVR XMEGA and the 32-bit Arm Cortex-M4 platforms under the POI-Selected scenario. For binary metrics, the optimal aggregation

operator determined in Section 5.2 is used.

5.3.1 AVR XMEGA

The convergence curves for the AVR XMEGA platform are shown in Figure 3.



(a) From 1 to 30 traces, with `averaging` = 100 (b) From 30 to 100 traces, with `averaging` = 100

Figure 3: Average Key Rank convergence curves for AES-128 on the AVR XMEGA platform, with the POI-Selected scenario. All metrics converge to 1 for 1,000 traces except `MoMV`

On the 8-bit AVR XMEGA target, physical power leakage is clean and exhibits a high Signal-to-Noise Ratio (SNR). Furthermore, the power consumption is highly linear and conforms closely to the first-order Hamming Weight model of the S-Box output.

In this linear environment, Pearson correlation (`P-corr`) is the most sample-efficient distinguisher, reaching an average key rank of 1.0000 across all 16 key bytes after **54 traces** (and reaching a key rank below 1.05 at 40 traces). This is consistent with side-channel theory: under a strictly linear leakage model with additive Gaussian noise, Pearson correlation is the optimal statistical distinguisher.

Multi-class metrics converge shortly after: the parametric ANOVA F-statistic (`F-stat`) and the non-parametric Kruskal-Wallis (`KW-stat`) statistic achieve full convergence at **70 traces** and **81 traces** respectively (with `VoM` and `MMoV` converging at **69** and **79 traces**). While these metrics do not presume a linear leakage profile, they evaluate variance across 9 distinct classes simultaneously, incurring a slight sample-size overhead to accurately resolve all empirical distributions.

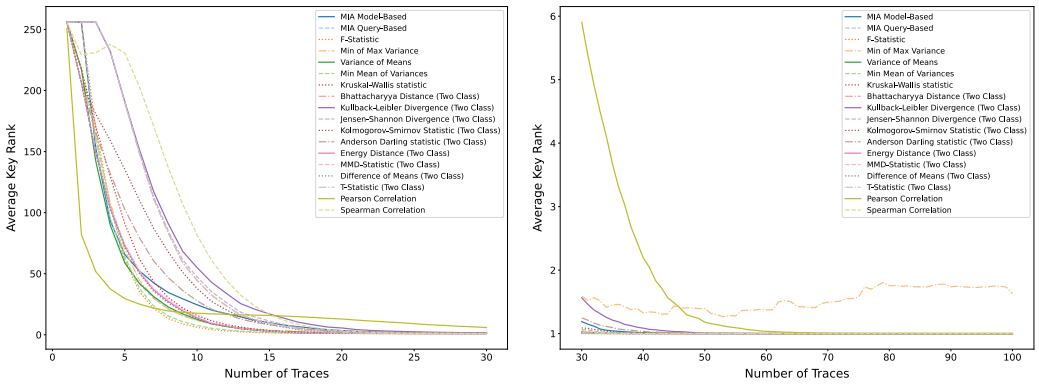
When combined with the Two-Class Bipartition (`TCB`) operator, binary metrics exhibit remarkable efficiency, dramatically outperforming their pairwise aggregation counterparts. In the very early-trace regime ($N \in [10, 30]$), `DoM` and `E-dist` converge even faster than `P-corr` (yielding average key ranks of 3.23 and 3.37 at $N = 20$, compared to 5.15 for `P-corr`). Full key recovery is achieved at **74 traces** for Difference of Means (`DoM`), Energy Distance (`E-dist`), and Bhattacharyya distance (`B-dist`), followed by Jensen-Shannon divergence (`JS-div`) at **78 traces**, Welch's t-statistic (`T-stat`) at **80 traces**, and `MMD-stat` at **84 traces**.

Furthermore, non-parametric tests based on empirical CDFs, namely Anderson-Darling (`AD-stat`) and Kolmogorov-Smirnov (`KS-stat`), which require over 300 traces under pairwise aggregation, achieve full key recovery at only **86 traces** and **91 traces** respectively, thanks to the substantial sample counts provided by macro-class pooling. Query-based MIA (`MIA-qb`) converges at **99 traces**, while model-based MIA (`MIA-mb`) and `KL-div` reach key

ranks of 1.0050 and 1.0006 at $N = 100$. Only `MoMV` fails to converge, plateauing at an average key rank of 13.47.

5.3.2 Arm Cortex-M4

The convergence curves for the Arm Cortex-M4 platform are shown in Figure 4.



(a) From 1 to 30 traces, with `averaging` = 100 (b) From 30 to 100 traces, with `averaging` = 100

Figure 4: Average Key Rank convergence curves for AES-128 on the Arm Cortex-M4 platform, with the POI-Selected scenario.

The 32-bit Arm Cortex-M4 architecture features pipelined execution, 32-bit internal buses, and parallel word manipulations. These characteristics generate substantial algorithmic noise and, more critically, prominent non-linear leakage components that deviate significantly from a linear Hamming Weight model. In this environment, we observe a paradigm shift in distinguisher efficacy.

Pearson correlation (`P-corr`) suffers severe performance degradation: its linearity assumption treats non-linear leakage components as additive noise, requiring **85 traces** to achieve an average key rank of 1.0000 (and only reaching KR ≤ 1.05 at 59 traces). At $N = 20$ traces, `P-corr` lags at an average key rank of 12.94.

In sharp contrast, multi-class metrics that make no assumptions regarding model linearity drastically outperform classical correlation. The parametric ANOVA F-statistic (`F-stat`) achieves full key recovery in only **29 traces** (reaching KR ≤ 1.05 at $N = 20$), representing an almost $3\times$ increase in sample efficiency over Pearson correlation. Similarly, Min Mean of Variances (`MMoV`) converges at **32 traces**, Kruskal-Wallis (`KW-stat`) at **38 traces**, and Query-based MIA (`MIA-qb`) and Spearman correlation (`S-corr`) at **43 traces**.

Binary metrics equipped with `TCB` also significantly outperform Pearson correlation: Energy Distance (`E-dist`) achieves full recovery at **38 traces**, Difference of Means (`DoM`) at **39 traces**, Welch's t-statistic (`T-stat`) and `MMD-stat` at **40 traces**, and Bhattacharyya distance (`B-dist`) at **41 traces** (all reaching KR ≤ 1.05 in only 24 to 25 traces). Non-parametric divergence and CDF metrics follow closely, with `JS-div` converging at **49 traces**, `KS-stat` at **57 traces**, and `KL-div` at **67 traces**.

These results clearly demonstrate that when evaluating modern 32-bit microcontrollers with non-linear leakage behaviors, distribution-based statistical distinguishers, whether multi-class or binary combined with Two-Class Bipartition, offer substantially greater resilience and sample efficiency than traditional correlation attacks.

5.3.3 Conclusion

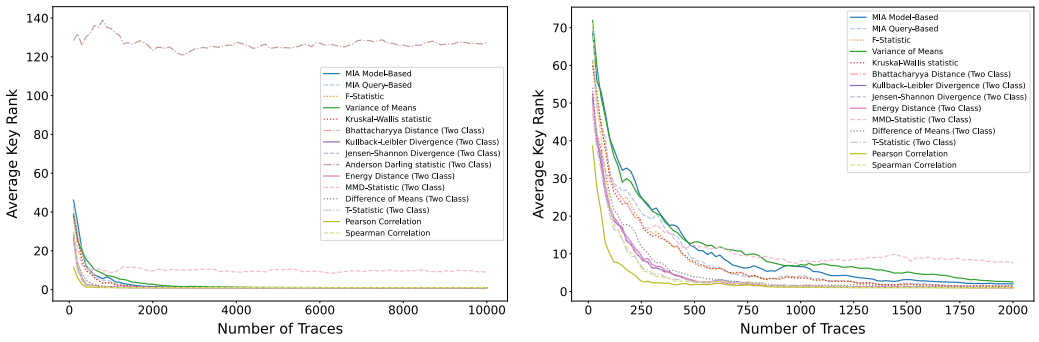
These experimental evaluations on unprotected AES-128 confirm the overall validity of our unified framework while highlighting the critical influence of the target architecture. On the 8-bit AVR XMEGA, where physical leakage adheres closely to a linear Hamming Weight model, classical Pearson correlation remains the most sample-efficient distinguisher. Conversely, on the 32-bit Arm Cortex-M4, prominent non-linear leakage components penalize linear correlation, allowing generic distribution-based metrics—both multi-class statistics and binary tests equipped with Two-Class Bipartition—to achieve substantially faster key recovery. Across both platforms, 17 out of the 18 evaluated metrics successfully achieve complete key recovery, with only `MoMV` failing to fully converge due to the noise sensitivity of its worst-case variance criterion.

5.4 Key Recovery Performance on Masked AES-128

We now evaluate the performance of the statistical metrics on the first-order software Boolean masked implementation of AES-128 (Herbst scheme [HOM06]). We target the same hardware platforms over $N = 10,000$ traces, this time with the POI-Auto scenario, since we cannot know *a priori* the leakage locations in the trace (for `MMD-stat`, the POI identified for `E-dist` were used, since it is too costly for being computed on all samples). Once again, for binary metrics, only the optimal aggregation operator `TCB` is used.

5.4.1 AVR XMEGA

The convergence curves for the AVR XMEGA platform are shown in Figure 5.



(a) From 1 to 10,000 traces, with `averaging` = 10, excluding `MoMV`, `MMoV` and `KS-stat`

(b) From 1 to 2,000 traces, with `averaging` = 10, excluding `MoMV`, `MMoV`, `KS-stat` and `AD-stat`

Figure 5: Average Key Rank convergence curves for Masked-AES-128 on the AVR XMEGA platform, with the POI-Auto scenario.

On the AVR XMEGA platform, we observe that almost all evaluated distinguishers successfully achieve complete key recovery, converging to the ideal average key rank of 1.00 at $N = 10,000$ traces. As shown in Figure 5b, classical Pearson correlation (`P-corr`) exhibits the highest sample efficiency, reaching KR ≤ 5 within only 200 traces, KR ≤ 2 at 400 traces, and converging to KR ≈ 1.01 at 2,000 traces. Non-parametric Energy Distance (`E-dist`) and Spearman correlation (`S-corr`) follow closely, achieving KR ≤ 2 at 680 and 780 traces respectively. Binary information divergences (`B-dist`, `KL-div`, `JS-div`) as well as parametric tests (`T-stat`, `DoM`) equipped with the `TCB` operator also demonstrate robust convergence, crossing KR ≤ 2 around 800 traces. Global multi-class metrics (`VoM`, `F-stat`, `KW-stat`, and `MIA`) converge slightly more slowly due to their higher partition dimensionality, achieving KR ≤ 2 between 1,360 and 1,920 traces before stabilizing at rank 1.00. Maximum

Mean Discrepancy (`MMD-stat`), evaluated using the POIs identified by `E-dist`, reaches an average key rank of 7.58 at 2,000 traces and 8.97 at 10,000 traces.

There are, however, two notable categories of exceptions that fail to recover the key. First, three metrics remain systematically locked at the worst-case rank of 256.0: `MoMV`, `MMoV`, and Kolmogorov-Smirnov (`KS-stat`). This complete degradation stems from the automated POI selection latching onto inactive or saturated trace samples exhibiting zero empirical variance. For variance minimization metrics (`MMoV`, `MoMV`), an invariant sample yields zero variance, representing an artificial global minimum; for `KS-stat`, degenerate identical sample distributions produce identical step CDFs across all classes. Consequently, at this selected sample point, all 256 subkey hypotheses produce completely identical scores. Under standard ranking evaluation, such a 256-way tie unfavorably assigns the worst possible rank of 256 to every candidate, including the correct key. Second, Anderson-Darling (`AD-stat`) fails through a distinct mechanism, stagnating around the random guessing baseline (KR ≈ 127.4). While its continuous integral formulation avoids discrete tie-breaking penalties, `AD-stat` heavily weights distribution tails via $(H(x)(1-H(x)))^{-1}$, severely de-emphasizing the central region of the power distribution where the physical leakage signals actually reside. Under noisy masked conditions, this tail sensitivity prevents `AD-stat` from distinguishing meaningful distribution shifts, rendering it ineffective for key recovery.

Finally, the fact that first-order distinguishers such as CPA (`P-corr`) and ANOVA (`F-stat`) completely recover the secret key on this masked implementation confirms the presence of micro-architectural hardware leakages. The 8-bit AVR XMEGA target probably exhibits bus transition glitches and register overwrite effects that recombine the secret mask and the masked intermediate state in hardware, re-introducing prominent first-order unmasked leakages into the power consumption.

5.4.2 Arm Cortex-M4

The convergence curves for the Arm Cortex-M4 platform are shown in Figure 6.

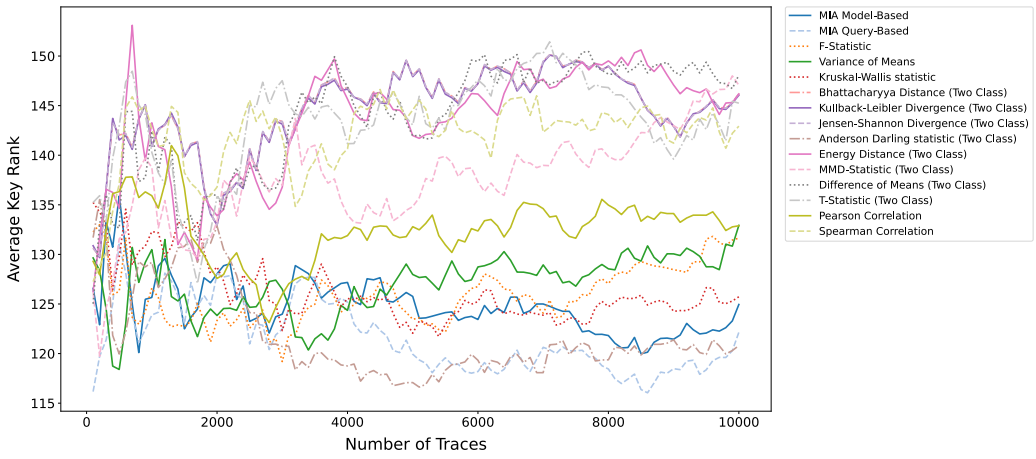


Figure 6: Average Key Rank convergence curves for Masked-AES-128 on the Arm Cortex-M4 platform, with the POI-Auto scenario, excluding `MoMV`, `MMoV` and `KS-stat`.

In contrast to the AVR XMEGA, the software implementation on this 32-bit platform maintains mask independence without compiler-induced unmasking or micro-architectural secret recombining. As a result, the first-order Boolean masking scheme remains fully sound and effective. We observe a universal failure of key recovery: all 18 evaluated metrics fail completely to extract the secret key. Their average key ranks fluctuate between 120

and 148, perfectly consistent with the random guessing baseline of 128 among 256 subkey candidates. Furthermore, `KS-stat`, `MMoV`, and `MoMV` remain locked at the worst-case rank of 256.0, suffering from the same tie-breaking degradation observed on the AVR XMEGA.

5.4.3 Conclusion

In summary, our empirical evaluation on masked implementations illustrates how statistical distinguishers exhibit contrasting behaviors depending on the leakage scenario. While hardware micro-architectural effects (such as bus transitions or overwrites) can re-introduce first-order univariate leakages on certain targets like the AVR XMEGA, the evaluated metrics display notably distinct sample efficiencies: correlation and parametric tests converge rapidly, non-parametric distance and divergence metrics follow closely, whereas global multi-class statistics require larger trace volumes to resolve key hypotheses. In contrast, metrics relying on tail-weighted empirical CDFs (`AD-stat`) or variance minimization (`MMoV`, `MoMV`) prove vulnerable to noise sensitivity and unmodulated sample ties. On other platforms like the Cortex-M4, all univariate evaluators fail as expected (KR ≈ 128). Because univariate analysis can only capture masked leakages when shares are physically recombined at a single time sample by hardware transitions or glitches, extending this unified framework to multivariate and second-order distinguishers constitutes a promising avenue for future work to evaluate their comparative efficacy when shares remain decoupled across multiple time samples.

5.5 Computational Cost

This section compares the theoretical computational complexity of each metric and the empirical execution times obtained across typical trace volumes. Table 2 summarizes the asymptotic complexity per sample point along with empirical wall-clock execution times measured across $N \in \{1\,000, 10\,000, 100\,000\}$ traces (benchmarked in single-thread mode to eliminate thread-scheduling artifacts). For binary metrics, the complexity reflects the Two-Class Bipartition (`TCB`) operator, which evaluates a single statistical test between the two macro-classes rather than $\binom{C+1}{2} = 36$ individual pairwise comparisons. All metric evaluations scale linearly with the number of subkey hypotheses $|\mathcal{K}| = 256$, which is omitted as an invariant constant factor $\mathcal{O}(1)$ in the asymptotic complexity column.

As shown in Table 2, parametric and linear metrics (`P-corr`, `T-stat`, `DoM`, `F-stat`, and variance-based metrics) exhibit linear scaling $\mathcal{O}(N)$, executing in less than 0.15 s per sample point across all 256 subkey hypotheses even at $N = 100\,000$ traces, with an exact $10\times$ increase per decade. Pairwise information divergences (`B-dist`, `KL-div`, `JS-div`) and mutual information metrics (`MIA-mb`, `MIA-qb`) incur a negligible binning constant ($\mathcal{O}(N+B)$ or $\mathcal{O}(N + C \cdot B)$), requiring only ~ 0.33 s at $N = 100\,000$ traces under the two-class bipartition. Non-parametric CDF tests (`KS-stat`, `AD-stat`) and the univariate Energy Distance (`E-dist`) require sorting sample traces, yielding $\mathcal{O}(N \log N)$ complexity; in the one-dimensional setting, `E-dist` avoids pairwise interactions via prefix sums over sorted observations, scaling at ~ 2.2 s for 100 000 traces alongside `KS-stat` (2.17 s) and `AD-stat` (3.21 s). In sharp contrast, `MMD-stat` exhibits strict quadratic growth $\mathcal{O}(N^2)$ due to all-to-all Gaussian kernel computations. Its empirical runtime scales by approximately $100\times$ for each $10\times$ increase in trace volume, rising from 0.84 s at $N = 1\,000$ to 84.16 s at $N = 10\,000$, and reaching 8 281.99 s (over 2.3 hours) per sample point at $N = 100\,000$ traces. This demonstrates the prohibitive quadratic computational wall of full-sample MMD evaluations compared to linear and sorting-based alternatives.

Table 2: Theoretical computational complexity and empirical wall-clock execution times (in seconds) per sample point across $N \in \{1\,000, 10\,000, 100\,000\}$ traces ($C + 1 = 9$ classes, histogram bins B, single-thread execution).

Family	Metric	Complexity	$N = 1\,000$	$N = 10\,000$	$N = 100\,000$
Dependence	`P-corr`	$\mathcal{O}(N)$	0.0010	0.0102	0.0999
	`S-corr`	$\mathcal{O}(N \log N)$	0.0008	0.0079	0.0878
Multi-Class	`F-stat`	$\mathcal{O}(N + C)$	0.0015	0.0130	0.1304
	`VoM`	$\mathcal{O}(N + C)$	0.0010	0.0086	0.0854
	`MMoV`	$\mathcal{O}(N + C)$	0.0009	0.0091	0.0917
	`MoMV`	$\mathcal{O}(N + C)$	0.0009	0.0090	0.0907
	`KW-stat`	$\mathcal{O}(N \log N + C)$	0.0008	0.0086	0.0890
	`MIA-mb`	$\mathcal{O}(N + C \cdot B)$	0.0014	0.0129	0.1455
	`MIA-qb`	$\mathcal{O}(N + C \cdot B)$	0.0010	0.0099	0.1113
Binary (Parametric)	`DoM`	$\mathcal{O}(N)$	0.0006	0.0063	0.0666
	`T-stat`	$\mathcal{O}(N)$	0.0009	0.0089	0.0872
Binary (Divergence)	`B-dist`	$\mathcal{O}(N + B)$	0.0033	0.0328	0.3307
	`KL-div`	$\mathcal{O}(N + B)$	0.0036	0.0357	0.3607
	`JS-div`	$\mathcal{O}(N + B)$	0.0033	0.0325	0.3290
Binary (Non-Param.)	`KS-stat`	$\mathcal{O}(N \log N)$	0.0172	0.1935	2.1698
	`AD-stat`	$\mathcal{O}(N \log N)$	0.0240	0.2798	3.2116
	`E-dist`	$\mathcal{O}(N \log N)$	0.0188	0.2017	2.2294
	`MMD-stat`	$\mathcal{O}(N^2)$	0.8371	84.165	**8282.0**

5.6 Discussion

The empirical evaluation of our unified statistical framework yields several key insights into the trade-offs between statistical assumptions, sample efficiency, and computational cost:

First-Order Means vs. Sample Variance Estimators. In low-noise or limited-trace regimes, Difference of Means (`DoM`) often matches or slightly outperforms Welch's t-statistic (`T-stat`). Welch's t-statistic incorporates sample variance estimators $\hat{\sigma}_i^2/n_i$ in its denominator, which can exhibit high estimation variance when class counts n_i are small, injecting noise into the distinguisher. In contrast, `DoM` relies purely on first-order sample means $\hat{\mu}_i$, offering enhanced stability. Computationally, Table 2 confirms that `DoM` is also faster (0.067 s vs. 0.087 s at $N = 100\,000$) by bypassing second-order sample variance normalizations.

Divergence Among Non-Parametric Metrics. Non-parametric metrics eliminate distribution shape assumptions, making them theoretically appealing for non-linear and masked leakages. However, our results demonstrate marked divergences among them. While 1D Energy Distance (`E-dist`) achieves full key recovery on both unprotected platforms and on the masked AVR target with efficient $\mathcal{O}(N \log N)$ scaling (2.2 s at $N = 100\,000$), `MMD-stat` is constrained by a severe quadratic computational wall ($\mathcal{O}(N^2)$, over 2.3 hours per sample point). Meanwhile, empirical CDF metrics reveal practical limitations: `AD-stat` fails to converge on masked AES (KR ≈ 127.4) because its tail weighting de-emphasizes the central distribution where leakages concentrate, whereas `KS-stat` is vulnerable to automated POI selection locking onto saturated zero-variance samples, causing 256-way tie-breaking failures (KR = 256).

Global Multi-Class Partitioning vs. Two-Class Divergences. Our benchmark highlights the practical synergy between partition scope and aggregation operators. Under the Two-Class Bipartition (`TCB`), pairwise information divergences (`B-dist`, `KL-div`, `JS-div`)

scale quasi-linearly ($\sim 0.33\,\mathrm{s}$ at $N = 100\,000$) and achieve fast convergence on masked AVR (KR ≤ 2 at ~ 800 traces). Conversely, global multi-class metrics (`F-stat`, `VoM`, `MIA`) operate in pure linear time ($\sim 0.09 - 0.15\,\mathrm{s}$) and excel on platforms with non-linear leakage components (Cortex-M4), while still reaching full key recovery on the masked target. In particular, ANOVA `F-stat` offers an optimal balance: it achieves complete key recovery across all successful setups while maintaining the minimal linear execution time of first-order CPA.

These findings show that distinguisher design is not about discovering isolated metrics, but navigating structured trade-offs across partition granularity, statistical moments, and aggregation operators to best match target leakage physics and computational constraints.

6 Related Work

The formalization of side-channel analysis (SCA) and the comparison of statistical key distinguishers have been active research topics over the past two decades.

Evaluation Frameworks in Side-Channel Analysis. Standaert et al. [SMY09] introduced a landmark unified framework for side-channel security evaluations. Their approach focuses on information-theoretic metrics (e.g., mutual information) and operational security metrics (e.g., success rate, guessing entropy, and key rank) to evaluate implementation security independently of specific attacks. While their framework provides a rigorous foundation for measuring leakage bounds and empirical attack success, it treats key distinguishers as atomic, monolithic functions. In contrast, our framework opens the algorithmic "black box" of distinguishers, factorizing their internal construction into subkey-induced leakage partitioning, statistical comparison functions, pairwise aggregation operators, and global decision rules. Prouff and Rivain [PR09] analyzed the mathematical foundations of Mutual Information Analysis (MIA) and established its asymptotic connection to CPA and DPA under Gaussian noise assumptions. Batina et al. [BGP+11] later provided a comprehensive experimental study of MIA, exploring the impact of probability density estimation methods, number of histogram bins, and sample sizes on its effectiveness as a side-channel distinguisher. Whitnall et al. [WOS11] systematically evaluated information-theoretic metrics across diverse noise regimes, highlighting parameter sensitivity in density estimation. Heuser et al. [HRG09] framed side-channel key recovery as a statistical hypothesis testing problem using goodness-of-fit metrics against theoretical leakage distributions. However, their model primarily addresses distribution fitting against expected templates rather than multi-class partition aggregation over non-parametric distance statistics.

Leakage Assessment vs. Key Recovery. A traditional division in SCA separates leakage assessment diagnostics from key recovery attacks. The Test Vector Leakage Assessment (TVLA) methodology, popularized by Goodwill et al. [GJJR11] and Schneider et al. [SM15], employs Welch's t-test to detect physical leakages by comparing two sets of traces (e.g., fixed vs. random inputs). Moradi et al. [MS18] generalized TVLA to multi-class partitions using the ANOVA F-test. Mather et al. [MOBW13] introduced an a priori statistical power analysis to formally evaluate the detection capabilities of leakage assessment tests, providing a principled basis for comparing their sensitivity. Whitnall and Oswald [WO19] critically analyzed the ISO 17825 standard for non-invasive attack testing, identifying fundamental methodological limitations in standardized leakage evaluation procedures. While leakage assessment tests are typically restricted to binary pass/fail hardware verification, our work demonstrates that any statistical leakage test—whether parametric (t-test, F-test) or non-parametric (Kolmogorov-Smirnov, Anderson-Darling, MMD, Energy Distance)—can be systematically converted into a full key recovery distinguisher. By formalizing symmetric and monotonic aggregation operators (such as Weighted Mean), our framework bridges the gap between two-class leakage detection tests and multi-class key recovery distinguishers.

Non-Parametric and Higher-Order Distinguishers. To overcome the limitations of linear power models, non-parametric metrics such as MIA [GBSV08], Maximum Mean Discrepancy (MMD), and Energy Distance have been proposed. Ding et al. [DZFL12] developed a statistical model for DPA that introduces an algorithmic confusion analysis, formalizing the relationship between key distinguisher success and the algebraic properties of the targeted cryptographic function. Durvaux and Standaert [DS16] proposed improved leakage detection techniques and extended them to the systematic identification of POIs in leakage traces, which is directly connected to our POI-Selected and POI-Auto experimental scenarios. Bronchain and Standaert [BS20] investigated the practical limits of closed-source security evaluations by dissecting side-channel countermeasures through comprehensive statistical assessments, reinforcing the need for systematic and reproducible evaluation frameworks. Non-parametric distance metrics do not assume a specific probability distribution shape, making them suitable for detecting higher-order leakages in masked implementations. Our taxonomy unifies these disparate tools under a single mathematical architecture, offering clear guidelines on when multi-class metrics (ANOVA F-test, MIA) or aggregated binary distance metrics (`T-stat`, `MMD-stat`, `E-dist`) should be preferred based on hardware target architecture, signal-to-noise ratio, and leakage model linearity.

Positioning of this Work. To the best of our knowledge, no prior study has established such a unified mathematical formalization that simultaneously deconstructs statistical distinguishers into orthogonal constituents, demonstrates how arbitrary statistical distances can be operationalized into key recovery attacks, and delivers an extensive empirical and computational benchmark across 18 metrics on both 8-bit and 32-bit architectures.

7 Conclusion

In this paper, we presented a generic statistical framework that unifies side-channel leakage assessment methodologies and key recovery distinguishers. The core contribution of this work lies in a fundamental shift of perspective: rather than treating correlation power analysis, mutual information analysis, Welch's t-tests, ANOVA F-tests, and distribution distance statistics as independent algorithms, we demonstrate that all statistical side-channel key distinguishers factorize into four orthogonal building blocks: a subkey-induced leakage partition; a core statistical comparison function applied to the partition (either binary pairwise or multi-class global); an aggregation operator combining pairwise binary distribution comparison matrices into a single scalar score; and a global decision rule maximizing or minimizing the resulting score across all key candidates.

We implemented 18 statistical metrics and 5 aggregation operators within this framework—provided as open-source software—and evaluated their performance across two target platforms (8-bit AVR XMEGA and 32-bit Arm Cortex-M4 microcontrollers) processing unprotected and first-order masked software AES implementations. The experimental results demonstrate the relevance of this modular approach: 17 out of 18 evaluated metrics achieve full key recovery on both platforms for unprotected AES, while on first-order masked AES, hardware micro-architectural leakages enable 13 metrics to recover the key on the 8-bit target whereas all univariate metrics fail on the 32-bit architecture as expected. Crucially, no single metric outperforms all others across all contexts. These findings empower security evaluators to design novel custom side-channel distinguishers by plugging modern statistical distance metrics or specialized aggregation operators into the evaluation pipeline. Future work includes extending this framework to multivariate spatio-temporal leakage distributions, deep learning-assisted partition models, and higher-order masked implementations.

References

[BCO04] Eric Brier, Christophe Clavier, and Francis Olivier. Correlation power analysis with a leakage model. In *CHES 2004*, volume 3156 of *LNCS*, pages 16–29. Springer, 2004.

[BGP+11] Lejla Batina, Benedikt Gierlichs, Emmanuel Prouff, Matthieu Rivain, François-Xavier Standaert, and Nicolas Veyrat-Charvillon. Mutual information analysis: a comprehensive study. *Journal of Cryptology*, 24(2):269–291, 2011.

[BS20] Olivier Bronchain and François-Xavier Standaert. Side-channel countermeasures' dissection and the limits of closed source security evaluations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(2):1–25, 2020.

[CAB25] Adrien Cassagne, Noé Amiot, and Manuel Bouyer. Dalek: An Unconventional and Energy-Aware Heterogeneous Cluster. working paper or preprint, August 2025.

[CRR02] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template attacks. In *CHES 2002*, volume 2523 of *LNCS*, pages 13–28. Springer, 2002.

[DS16] François Durvaux and François-Xavier Standaert. From improved leakage detection to the detection of points of interests in leakage traces. In *EUROCRYPT 2016*, volume 9665 of *LNCS*, pages 240–262. Springer, 2016.

[DZFL12] A. Adam Ding, Liwei Zhang, Yunsi Fei, and Pei Luo. A statistical model for DPA with novel algorithmic confusion analysis. In *CHES 2012*, volume 7428 of *LNCS*, pages 233–250. Springer, 2012.

[GBR+12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[GBSV08] Benedikt Gierlichs, Lejla Batina, Preeti Subramanian, and Ingrid Verbauwhede. Mutual information analysis. In *CHES 2008*, volume 5154 of *LNCS*, pages 426–442. Springer, 2008.

[GJJR11] Gilbert Goodwill, Benjamin Jun, Josh Jaffe, and Pankaj Rohatgi. A testing methodology for side-channel resistance validation. *NIST Non-Invasive Attack Testing Workshop (NIAT)*, 2011.

[HOM06] Christoph Herbst, Elisabeth Oswald, and Stefan Mangard. An aes smart card implementation resistant to power analysis attacks. In *ACNS*, volume 3989, pages 239–252. Springer, 2006.

[HRG09] Annelie Heuser, Olivier Rioul, and Sylvain Guilley. Goodness-of-fit side-channel distinguishers. In *CHES 2014*, volume 8731 of *LNCS*, pages 89–107. Springer, 2009.

[KJJ99] Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. Differential power analysis. In *CRYPTO '99*, volume 1666 of *LNCS*, pages 388–397. Springer, 1999.

[MOBW13] Luke Mather, Elisabeth Oswald, Joe Bandenburg, and Marcin Wójcik. Does my device leak information? an a priori statistical power analysis of leakage detection tests. In *ASIACRYPT 2013*, volume 8269 of *LNCS*, pages 486–505. Springer, 2013.

[MOP07] Stefan Mangard, Elisabeth Oswald, and Thomas Popp. Power analysis attacks: Revealing the secrets of smart cards. Springer, 2007.

[MS18] Amir Moradi and Tobias Schneider. Side-channel leakage assessment: Towards validation an entire system. In *COSADE 2018*, volume 10815 of *LNCS*, pages 3–18. Springer, 2018.

[OC14] Colin O'Flynn and Zhizhang David Chen. Chipwhisperer: An open-source platform for hardware embedded security research. In *COSADE*, 2014.

[PR09] Emmanuel Prouff and Matthieu Rivain. Theoretical analysis of mutual information analysis and comparison with other side-channel distinguishers. In *EUROCRYPT 2009*, volume 5479 of *LNCS*, pages 638–653. Springer, 2009.

[SM15] Tobias Schneider and Amir Moradi. Leakage assessment methodology – a clear roadmap for side-channel evaluations. In *CHES 2015*, volume 9293 of *LNCS*, pages 495–513. Springer, 2015.

[SMY09] François-Xavier Standaert, Tal G. Malkin, and Moti Yung. A unified framework for the analysis of side-channel key recovery attacks. In *EUROCRYPT 2009*, volume 5479 of *LNCS*, pages 443–461. Springer, 2009.

[SR04] Gábor J. Székely and Márton L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5:1–6, 2004.

[WO19] Carolyn Whitnall and Elisabeth Oswald. A critical analysis of ISO 17825 ('testing methods for the mitigation of non-invasive attack classes against cryptographic modules'). In *ASIACRYPT 2019*, volume 11923 of *LNCS*, pages 256–284. Springer, 2019.

[WOS11] Carolyn Whitnall, Elisabeth Oswald, and Francois-Xavier Standaert. The suitability of information theoretic metrics for side-channel analysis. In *CHES 2011*, volume 6917 of *LNCS*, pages 49–64. Springer, 2011.