Bounded Information: PAC Certification of Multivariate Side-Channel Traces

Kuheli Pratihar[0000−0003−4486−4903], Nimish Mishra[0000−0002−8585−9425], and Debdeep Mukhopadhyay[0000−0002−6499−8346]

Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India
{its.kuheli96,neelam.nimish,debdeep.mukhopadhyay}@gmail.com

Abstract. Side-channel leakage certification aims to quantify what an attacker can learn about a secret variable from observed leakage. Existing information-theoretic estimators, such as perceived information (PI), hypothetical information (HI), and nonparametric mutual information (MI), aim to quantify distributional leakage, but they become unstable in high-dimensional traces and do not provide a finite-sample certificate of the best attacker. We introduce *Bounded Information* (BI), a probably approximately correct (PAC)-style finite-sample certification method that upper-bounds the exact-recovery success of a fixed attacker scope on unseen traces. The attacker suite certificate, $\mathrm{BI}^{\mathrm{suite}}$, applies a KL-binomial confidence interval with a union bound over a fixed suite that contains every trained attacker, preprocessing choice, and hyperparameter. BI therefore turns standard profiled-attack evaluation into an auditable certificate with an explicit attacker scope and confidence level. We further define $\mathrm{BI}^{\mathrm{loc}}(\varepsilon)$ to bound the recovery success of attackers whose normalized scores differ by at most ε from those of a model in the suite. Across eight side-channel benchmarks with trace dimensions up to 7,000, BI provides stable certificates without density estimation, and its tightness diagnostics tell an evaluator whether a certified value is a genuine measurement of leakage or a conservative bound that more attack traces would tighten.

Keywords: Side-channel analysis · Leakage certification · PAC bounds · Multivariate leakage · Deep learning side-channel attacks.

1 Introduction

Side-channel analysis (SCA) threatens cryptographic implementations by exploiting physical observables such as power consumption or electromagnetic leakage [36]. These threats also extend beyond cryptographic implementations, enabling neural-network architecture extraction [2], GPU-based website fingerprinting [32], and recovery of graph structure and language-model inputs and outputs through CPU-to-GPU cache attacks on Apple Silicon [42]. Formal abstractions in the probing [22] and noisy-leakage [35,15,16] settings provide conceptual targets for what a secure implementation should look like, but verifying

whether a real device meets those targets is a separate task known as *leakage certification*. Given a finite set of traces measured from a device-under-test (DUT), the certifier never has direct access to the true leakage distribution. The certification task is instead to reason about how closely a practical adversary can approximate it. Therefore, leakage certification aims to quantify the *gap* between an *ideal adversary* (who is given perfect knowledge of the actual leakage distribution of the DUT) and the *concrete adversary* (practical instantiation of the adversary under realistic circumstances).

The standard tools for this task, namely density-based estimators such as perceived information (PI), hypothetical information (HI), and nonparametric mutual information (MI), become unreliable as the trace dimension grows [11]. Protected implementations push exploitable leakage into the joint structure across many time samples [19,27,4], which low-dimensional projections may miss. To compound this, information-theoretic leakage numbers from the same data vary with estimation methodology, undermining their role as stable certificates. A further limitation is that these metrics characterize the *information content* of traces, whereas an evaluator ultimately needs to know what a class of attackers recovers from attack traces. Unlike finite-sample success certificates, density estimates are not parameterized by models that attackers use for SCA.

We address both gaps with *Bounded Information* (BI). For an adversary family, BI certifies an upper bound on the attacker's success rate on attack traces. BI adds two ingredients to this certified bound that prior work lacks. The first is a *certification methodology* that combines PAC-style confidence accounting with an explicit declaration of the adversary family. The second is a *capacity-aware* family-wide analysis that extends the certificate from the trained attackers to nearby neural score functions. Once the attacker scope is fixed, the resulting confidence accounting does not depend on the trace dimension. This combination provides an auditable route from a trained suite to nearby neural score functions.

1.1 Existing Estimators and Their Limitations

Prior work has developed information-theoretic metrics that characterize leakage through the true distribution, fitted probability models, or learned representations. Their different targets and estimation assumptions explain both their usefulness and the need for a separate confidence bound on recovery probability.

Mutual information and model-dependent metrics. A natural certification target is mutual information, defined as $I(X;Y) = H(X) - H(X \mid Y)$. It measures the reduction in Shannon entropy about the secret variable X after observing the leakage trace Y. Since evaluators do not know the true joint distribution $p_{\text{chip}}(x, y)$, they estimate information from finite data or use a fitted posterior model $q_\theta(x \mid y)$ [3]. PI and HI characterize information using such a model, but differ in the distribution under which they evaluate it.

Renauld et al. [37] introduced PI to evaluate a fitted posterior model under the true device distribution. It is defined as

$$\mathrm{PI}_\theta(X;Y) \;=\; H(X) \;+\; \mathbb{E}_{(X,Y)\sim p_{\text{chip}}}\left[\log_2 q_\theta(X \mid Y)\right]. \tag{1}$$

Table 1: The metrics differ in their objectives and evaluation assumptions. The last three columns indicate whether the stated construction gives a finite-sample recovery certificate, computes its statistical bound without an explicit trace-dimension term, and covers a specified attacker family. A dash indicates that the corresponding guarantee is not part of the construction summarized here.

Metric	Primary objective	Measured quantity & assumption	PAC bound	No d term	Family coverage
PI [37]	It lower-bounds MI.	It evaluates a fitted posterior under the device distribution.	—	—	—
EPI [24]	It calibrates PI.	It adjusts model probabilities to better reflect attack success.	—	—	—
TI [29]	It evaluates training information.	It evaluates the fitted model using its training data.	—	—	—
HI [8]	It evaluates model-based information.	Its empirical upper-bound interpretation requires the stated estimation assumptions.	—	—	—
LPI [25]	It measures latent information.	It characterizes information in a trained feature extractor's representation.	—	—	—
BI (this work)	It certifies recovery probability.	It bounds population recovery probability within a specified attacker family using independent attack traces.	✓	✓	✓

PI is a lower bound on MI and equals MI when q_θ matches the true posterior. Model mismatch can make PI negative even when the traces contain information about the secret variable. Bronchain et al. [8] introduced HI as a complementary quantity evaluated under the model-induced joint distribution p_θ, defined as

$$\mathrm{HI}_\theta(X;Y) \;=\; H(X) \;+\; \mathbb{E}_{(X,Y)\sim p_\theta}\left[\log_2 q_\theta(X \mid Y)\right]. \tag{2}$$

HI characterizes information under the fitted model, so its relation to the real-device MI depends on the relation between p_θ and p_{chip}. The empirical constructions also differ. Empirical PI averages $\log_2 q_\theta(X \mid Y)$ over independent attack traces, while histogram-based empirical HI uses frequencies of secret values and discretized leakage values in the profiling data. Under the assumptions of the discrete estimation setting, empirical PI and HI support lower- and upper-bound interpretations in expectation [8,11]. Such expectation statements do not by themselves provide confidence bounds for a particular finite dataset. When estimation uses discretized traces, the information target also concerns those discretized observations.

Probability calibration introduces another distinction between these metrics and recovery success. Models can assign different probabilities while predicting the same secret value on every trace. Ito et al. [24] introduced effective perceived information (EPI) to account for such variations and align the probability-based evaluation more closely with attack performance.

Further model-dependent diagnostics. Training information (TI), introduced by Masure et al. [29], evaluates a fitted model on its training data to

study finite-sample estimation bias and the limitations of the model class. Latent perceived information (LPI), introduced by Ito et al. [25], characterizes mutual information between a learned representation and the secret variable in deep-learning side-channel attacks (DL-SCA). This information supports an upper bound on the corresponding model's recovery success. These diagnostics complement PI and HI, but evaluating a fitted model or representation does not, by itself, certify the maximum recovery probability over a suite of attackers.

The curse of dimensionality. Probability calibration does not remove the difficulty of learning the relationship between a high-dimensional trace and the secret variable. Nonparametric k-nearest-neighbor estimators, such as GKOV [17], avoid histogram binning but still face estimation difficulties as trace dimension increases. Chowdhury et al. [11] document slow convergence and dimensionality limitations in multivariate information estimation. Histogram-based methods additionally replace continuous traces with discretized observations, which cannot contain more MI about the secret than the original traces. These limitations motivate recovery certification that avoids explicit density estimation.

Table 1 summarizes the resulting gap. Information estimation can become difficult at high trace dimension, and the estimated quantity need not directly bound the recovery success of a specified attacker family. Model-dependent diagnostics describe particular fitted models, while quantities such as maximal leakage [23] concern unrestricted inference rather than recovery within a fixed trained suite. Our target is a finite-sample confidence bound on the maximum population recovery probability over the evaluated family. Its statistical precision depends on the number of independent attack traces, which is distinct from the dimension of each trace.

Difference from prior work. BI uses empirical recovery success to compute a confidence bound for a specified attacker family. The finite-suite certificate, $\mathrm{BI}^{\mathrm{suite}}$, covers every model in a fixed trained suite, including models with different architectures. The local-family certificate, $\mathrm{BI}^{\mathrm{loc}}(\varepsilon)$, extends this guarantee through uniform constraints on normalized score differences from the trained reference models. These constructions convert attack evaluations into recovery guarantees with an explicit confidence level and family coverage. Under a uniform secret distribution, BI expresses the probability bound in bits relative to random guessing. Interpreting it as a bound on unrestricted recovery or mutual information additionally requires Bayes-completeness, meaning the family's supremum recovery probability is Bayes-optimal.

1.2 Technical Overview

The success probability to certify. BI asks how successfully an attacker in a specified family can recover the secret variable from one leakage trace. Let $\mathcal{H}$ be a family of classifiers mapping traces in $\mathcal{Y}$ to candidate secret values in $\mathcal{X}$. The family's supremum population recovery probability is

$$P^{\mathcal{H}}_{\mathrm{guess}}(X \mid Y) \;:=\; \sup_{h \in \mathcal{H}} \Pr[h(Y) = X], \tag{3}$$

where the probability is over a new secret value and trace drawn from the same joint distribution as the attack data. The attack set provides an empirical recovery probability for each evaluated model. Confidence bounds account for the statistical uncertainty between these empirical values and the population recovery probabilities.

Constructing the trained-suite certificate. The evaluator trains candidate models on profiling traces and uses validation traces to select the models and preprocessing. Once these choices are fixed, each attacker is evaluated on independent and identically distributed attack traces, producing one success or failure per trace. Binomial confidence bounds convert each model's success count into lower and upper bounds on its population recovery probability. Allocating failure probability $\delta/(2M)$ to each of the two bounds for each of the M models makes all bounds valid simultaneously with confidence at least $1-\delta$. The maximum of the lower bounds and the maximum of the upper bounds then bound the suite's maximum population recovery probability.

Expressing the suite's upper bound in bits gives $\mathrm{BI}^{\mathrm{suite}}$. For a uniform secret with K possible values, a certificate of b bits bounds recovery probability by $2^b/K$. Random guessing corresponds to zero bits, while $\log_2 K$ bits gives the trivial probability bound of one. Applying a one-sided binomial bound to a single model selected using validation traces gives the singleton certificate $\mathrm{BI}^{\mathrm{sing}}$.

Extending the certificate to nearby models. A neural network assigns a class score to each candidate secret value and predicts the class with the highest score, using a fixed rule to resolve ties. The classification margin is the normalized true-class score minus the highest normalized competing score. If each normalized class score changes by at most ε, this margin changes by at most 2ε. A trace with reference margin below -2ε therefore remains incorrectly classified throughout the score neighborhood. BI counts every attack trace whose reference margin is at least -2ε and applies a binomial upper confidence bound to this relaxed-margin count. The resulting $\mathrm{BI}^{\mathrm{loc}}(\varepsilon)$ covers every model in the neighborhood, including models that have not been trained or individually evaluated. The relaxed count can also include traces that no model in the neighborhood recovers, so the resulting upper bound need not be attained.

Establishing that a particular model belongs to a score neighborhood requires the score-distance bound to hold on every admissible trace. We give sufficient conditions for multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and Transformers by tracking how parameter perturbations affect output scores. For CNNs with shared filters and normalized average pooling, the score bound need not increase with trace dimension when the patch-domain and parameter bounds remain fixed. The MLP and Transformer analyses account for their input norms, layers, and, where relevant, token sequence length. These architectural calculations establish family membership, while the subsequent binomial confidence calculation has no explicit dependence on trace dimension.

Evaluating a radius grid fixed independently of the certification traces shows how the upper bound changes as the covered family expands. We allocate the failure probability across all reference models and grid values to obtain simul-

taneous coverage. At radius zero, the local and suite upper bounds agree when they use the same traces, equal success counts, and the same failure probability for each upper bound. Ties for the highest score can make the relaxed-margin count larger than the ordinary success count, as explained in Remark 3.

Contributions. The paper makes the following contributions.

- We construct a finite-sample confidence interval for the maximum population recovery probability over a fixed suite of trained attackers (Proposition 2). Reporting the empirical maximum together with the lower and upper confidence bounds distinguishes measured attack performance from statistical uncertainty.
- We extend certification beyond the trained suite through local score families and relaxed-margin counts (Theorem 1). Architecture-specific perturbation bounds give sufficient conditions for models to belong to these families.
- We identify conditions under which a family using shared local features and additive aggregation can approach the Bayes-optimal recovery probability, together with stronger margin conditions under which a member attains it (Proposition 4). Each local feature processes a temporal window of τ consecutive sampling positions. The result relates information available within these windows to the family's ability to represent an optimal recovery rule.
- We evaluate the method on eight benchmark settings. The experiments examine how the reported bounds depend on attack-set size, suite composition, and score radius. We also relate single-trace recovery to empirical key-recovery attacks that combine several traces. The CHES experiments include a trace-dimension sweep up to $d = 7{,}000$ and are reported as numerical diagnostics because their preprocessing uses attack-set statistics, leaving the required independence condition unestablished.

1.3 Scope and Positioning

BI bounds population recovery probability for a specified family of profiled attackers evaluated on independent labeled attack traces. A small suite upper bound limits the recovery probability of every suite member at the stated confidence level, while the local extension covers additional models satisfying the uniform score-distance condition. More independent attack traces can reduce statistical uncertainty for a fixed family. Covering stronger attackers requires expanding that family or establishing that its supremum recovery probability already matches the optimal value. Single-trace recovery also differs from key recovery that combines scores across several traces, for which Section 4.4 gives composition assumptions and a direct certification procedure.

When the certified family is Bayes-complete, its supremum recovery probability equals that of an optimal attacker with knowledge of the leakage distribution. The certificate then also bounds unrestricted guessing probability and, for a uniform secret, mutual information (Section 3.5). The result for additive local features gives sufficient conditions for this equality at the family level. It does not establish that a finite training run finds a Bayes-optimal model or that

a particular score neighborhood contains one. Comparisons between devices or countermeasures must therefore account for both the evaluation protocol and the attacker families covered by their bounds.

Organization. Section 2 introduces the background and notation. Section 3 develops the BI framework, and Section 4 gives the neural-model constructions and composition results. Section 5 presents the experiments, and Section 6 discusses limitations and future work. The appendices provide the proofs and additional experimental details.

2 Background and Certification Setting

Leakage certification uses a finite number of measured traces to assess what an attacker can recover about a secret variable. This section defines the measurement setting and attacker family and then explains how the observed attack success supports a confidence bound on recovery from new traces.

2.1 Setting and Threat Model

We consider a device that processes a secret variable X and produces a leakage trace $Y \in \mathbb{R}^d$, where d is the trace dimension. The variable X takes values in a finite set $\mathcal{X}$ of size K. For example, an AES S-box output takes one of 256 possible values. We denote the joint distribution of the secret variable and the measured trace by $p_{\text{chip}}(x, y)$, which depends on both the device and the measurement setup. Unless stated otherwise, X is uniform over $\mathcal{X}$, and information quantities are measured in bits.

The evaluator receives a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of independent and identically distributed (i.i.d.) trace-label pairs drawn from p_{chip}. Independence is assumed across device executions, although time samples within the same trace may be statistically dependent. An attacker can therefore exploit dependence among samples within a trace without violating the independence assumption used for certification.

An attacker predicts the secret variable from a trace using a function $h : \mathcal{Y} \to \mathcal{X}$, where $\mathcal{Y}$ is the trace domain. The certificate covers an attacker family $\mathcal{H}$, which is a hypothesis class. This family may consist of a single trained model, a finite suite of trained models, or a larger class subject to specified model constraints. After specifying $\mathcal{H}$ and a failure probability δ, the evaluator uses independent attack traces to bound the maximum recovery probability within this family with confidence at least $1 - \delta$.

The measurement setup determines the distribution being evaluated, whereas the choice of $\mathcal{H}$ determines which attackers the guarantee covers. Training affects the success of the models included in the evaluation, but a certificate for those models does not automatically cover stronger attackers outside the family. A small certified bound must therefore be interpreted together with the family for which it was computed.

Remark 1 (Certification scope). Our main setting is a profiled side-channel attack in which the attacker is trained using labeled profiling traces and evaluated using independent attack traces. Non-profiled attacks, such as DDLA [41], can also be considered when their training and evaluation satisfy this separation. The essential requirement is that the evaluated prediction rule be fixed independently of the attack traces used for certification. Section 4.4 discusses composition and direct certification of attacks that combine multiple traces for key recovery. Appendix B.2 further discusses the scope of the evaluation.

2.2 Certification Goal

Our goal is to bound the probability that an attacker in $\mathcal{H}$ correctly recovers X from a new trace drawn from the same distribution. Exact recovery requires predicting the correct value of X among its K possible values from a single trace. In our experiments, X is an 8-bit intermediate value with $K = 256$, so the observed recovery success is the model's classification accuracy.

Single-trace accuracy and key recovery from several traces answer different questions. An attack may combine probability scores across traces even when its prediction on an individual trace is often incorrect, so classification accuracy alone need not predict the number of traces required for key recovery [33]. The certificate below concerns recovery of the specified variable from one trace. Section 4.4 states the conditions required to bound key-recovery success when an attack combines evidence across multiple traces or target variables.

Definition 1 (PAC-style certificate). *Let $\mathcal{H}$ be an attacker family fixed independently of an attack set S_{te} containing n_{te} i.i.d. trace-label pairs drawn from p_{chip}. For $\delta \in (0, 1)$, a probably approximately correct (PAC)-style certificate is a value $\varepsilon^\star$ computed from S_{te}, $\mathcal{H}$, and δ such that, with probability at least $1 - \delta$ over the sampling of S_{te},*

$$P^{\mathcal{H}}_{\mathrm{guess}}(X \mid Y) \leq \frac{1}{K} + \varepsilon^\star, \tag{4}$$

where

$$P^{\mathcal{H}}_{\mathrm{guess}}(X \mid Y) := \sup_{h \in \mathcal{H}} \Pr[h(Y) = X] \tag{5}$$

is the supremum of the recovery probabilities over $\mathcal{H}$ for a new trace from the same distribution.

The term $1/K$ is the success probability of random guessing, and $\varepsilon^\star$ bounds the family's advantage above this baseline. The confidence level describes the reliability of the bound across repeated sampling of the attack set. It therefore refers to the probability that the bound holds, whereas the recovery probability describes the success of an attacker on an individual trace. BI expresses the resulting success bound in bits, as defined in Section 3.2.

The PAC interpretation concerns the generalization gap between empirical success on a finite attack set and population success under the same distribution. The guarantee applies to the specified attacker family without assuming that training finds an optimal model or that the family contains an optimal attacker.

2.3 From Attack Success to a Confidence Bound

We divide the available data into profiling traces for training, validation traces for model selection, and independent attack traces for certification. The models, preprocessing, and attacker family are fixed independently of the attack set. This separation ensures that the confidence bound applies to prediction rules selected without using the certification data. Section 3 describes this protocol.

For a fixed attacker, recovery on each attack trace is a Bernoulli trial whose outcome indicates whether the predicted secret value is correct. Under the sampling assumptions above, these trials are independent and have the same success probability. A binomial confidence bound therefore provides an upper bound on the population recovery probability from the observed success count. For a finite suite, a union bound makes the individual guarantees hold simultaneously over all its members, allowing the evaluator to bound the best attacker's success even when the strongest member is not known in advance [39,31].

More generally, a uniform bound on the generalization gap across an attacker family provides a certificate for that family. For uniform X, we define the family's recovery advantage as

$$\mathrm{Adv}^{\mathcal{H}}(X \mid Y) := \sup_{h \in \mathcal{H}} \left(\Pr[h(Y) = X] - \frac{1}{K} \right). \tag{6}$$

Replacing the population recovery probability with the empirical accuracy on the attack set gives the empirical recovery advantage

$$\widehat{\mathrm{Adv}}^{\mathcal{H}}(X \mid Y; S_{\mathrm{te}}) := \sup_{h \in \mathcal{H}} \left(\frac{1}{n_{\mathrm{te}}} \sum_{(x,y) \in S_{\mathrm{te}}} \mathbf{1}\{h(y) = x\} - \frac{1}{K} \right). \tag{7}$$

For a finite suite, this quantity is obtained by evaluating each model and taking the largest observed advantage.

Suppose a uniform deviation bound $\Gamma(\mathcal{H}, n_{\mathrm{te}}, \delta)$ bounds population success minus empirical success simultaneously over $\mathcal{H}$ with probability at least $1 - \delta$. Whenever this bound holds, the population recovery advantage satisfies

$$\mathrm{Adv}^{\mathcal{H}}(X \mid Y) \le \widehat{\mathrm{Adv}}^{\mathcal{H}}(X \mid Y; S_{\mathrm{te}}) + \Gamma(\mathcal{H}, n_{\mathrm{te}}, \delta) =: \varepsilon^{\star}. \tag{8}$$

This expression separates the observed recovery advantage from the statistical margin needed to obtain a confidence guarantee. The form of Γ depends on the attacker family and the chosen statistical bound, as formalized in Proposition 1.

Section 3 constructs the singleton and finite-suite certificates using binomial confidence bounds. The local-family extension in Section 4 instead constrains deviations in model output scores relative to fixed reference models. These constraints allow counts obtained from the reference models to bound recovery success uniformly over the local family. The resulting certificate therefore covers the larger family without evaluating each of its members. Table 2 summarizes the notation used in these constructions.

Table 2: Notation for the attacker family, attack traces, and confidence bounds.

Symbol	Meaning
$\mathcal{D},\ n$	$\mathcal{D}$ contains n trace-label pairs.
$S_{\mathrm{te}},\ n_{\mathrm{te}}$	S_{te} contains n_{te} independent attack traces used for certification.
$\mathcal{H}$	$\mathcal{H}$ is the attacker family covered by the certificate.
$\mathcal{H}_{\mathrm{tr}},\ M$	$\mathcal{H}_{\mathrm{tr}} = \{h_1, \ldots, h_M\}$ is a fixed suite of M trained attackers.
$\widehat{p}_j$	$\widehat{p}_j$ is the empirical recovery probability of attacker h_j on the attack set.
p_{obs}	$p_{\mathrm{obs}} = \max_j \widehat{p}_j$ is the maximum empirical recovery probability in the suite.
$P^{\mathcal{H}}_{\mathrm{guess}}(X \mid Y)$	This is the supremum of the recovery probabilities over $\mathcal{H}$ on a new trace.
$\mathrm{Adv}^{\mathcal{H}}(X \mid Y)$	This is the recovery advantage over random guessing, defined as $P^{\mathcal{H}}_{\mathrm{guess}}(X \mid Y) - 1/K$.
$\Gamma(\mathcal{H}, n_{\mathrm{te}}, \delta)$	Γ is a statistical deviation bound that holds simultaneously over the family.
$\varepsilon^\star$	$\varepsilon^\star$ is a certified upper bound on the family's recovery advantage.
δ	δ is the failure probability of the certificate, which holds with confidence at least $1 - \delta$.
$\beta(\varepsilon, K)$	$\beta(\varepsilon, K) = \log_2(1 + K\varepsilon)$ converts recovery advantage to bits, as defined in Section 3.2.

3 Bounded Information for Multivariate Traces

BI uses empirical recovery success on independent attack traces to bound the population recovery probability of a specified attacker family. This section explains how the approach applies to multivariate leakage and expresses the resulting probability bound in bits. It then constructs certificates for three attacker families and discusses their statistical uncertainty and the conditions required for a mutual information bound.

3.1 Certifying Recovery from Multivariate Traces

An attacker may exploit statistical dependence among time samples within a trace to recover the secret variable, particularly when protection mechanisms distribute leakage across time or across multiple shares. Selecting a few points of interest can discard information carried by this dependence. However, direct information estimation at high trace dimension can require more data than the evaluator has available [11].

BI addresses this problem by evaluating whether an attacker predicts the secret variable correctly. For a fixed attacker, each prediction produces a Bernoulli outcome regardless of the trace dimension. The outcomes on independent attack traces therefore support a binomial confidence bound on the attacker's population recovery probability. This bound depends on the empirical success count,

the attack-set size, and the confidence level, without an explicit dependence on trace dimension. The trace dimension can still affect recovery success and the computational cost of training and evaluation.

Following the setting in Section 2.1, we use profiling traces for training, validation traces for model selection, and independent attack traces for certification. The suite of evaluated attacks, including their preprocessing, is fixed before the attack traces are used. This separation allows the same principle used to evaluate one profiled attack [20] to support a simultaneous guarantee for all attacks in a fixed suite, as developed below.

3.2 The Bounded Information Certificate

The primary output of certification is an upper bound on the family's recovery probability. BI expresses this bound in bits, with random guessing corresponding to zero bits and perfect recovery corresponding to $\log_2 K$ bits for a secret alphabet of size K. The conversion provides a common scale for reporting recovery bounds. Its interpretation as a mutual information bound requires the additional condition stated in Section 3.5.

Definition 2 (Bounded Information (BI)). *Let X be uniform on an alphabet of size $K \geq 2$, and let $\mathcal{H}$ be an attacker family. For $0 \leq \varepsilon \leq 1 - 1/K$, define*

$$\beta(\varepsilon, K) := \log_2(1 + K\varepsilon). \tag{9}$$

The population Bounded Information *is*

$$\mathrm{BI}^{\mathcal{H}}_{\mathrm{pop}}(X;Y) := \log_2\big(\max\{1, KP^{\mathcal{H}}_{\mathrm{guess}}(X \mid Y)\}\big)\,. \tag{10}$$

Thus, recovery at or below chance is reported as zero bits. Given a dataset $\mathcal{D}$ and a PAC-style advantage bound $\varepsilon^\star$ satisfying Equation (4)*, define*

$$\overline{\varepsilon}^\star := \min\{1 - 1/K, \max\{0, \varepsilon^\star\}\}. \tag{11}$$

The BI certificate *is*

$$\mathrm{BI}(\mathcal{D}, \mathcal{H}, \delta) := \beta(\overline{\varepsilon}^\star, K). \tag{12}$$

based on Definition 1, this certificate bounds $\mathrm{BI}^{\mathcal{H}}_{\mathrm{pop}}(X;Y)$ from above.

The clipping in this definition keeps the reported value between zero and $\log_2 K$ without weakening the confidence guarantee. If the certified recovery bound is below the random-guessing baseline, replacing it with that baseline preserves the upper bound. Similarly, a bound above one can be replaced with one because recovery probability cannot exceed one. For bounds between these limits, substituting $P^{\mathcal{H}}_{\mathrm{guess}} \leq 1/K + \varepsilon^\star$ gives the expression $\log_2(1 + K\varepsilon^\star)$.

A BI certificate of b bits therefore bounds the family's recovery probability by $2^b/K$. For an 8-bit target, a 2-bit certificate bounds single-trace recovery success by $4/256$, or 1.5625%, at the specified confidence level. A zero-bit certificate bounds success by chance, while an 8-bit certificate gives the trivial upper bound

of one. These values are upper bounds on population recovery probability and need not be attained by an attacker.

Statistical uncertainty and the choice of attacker family determine how the certificate should be interpreted. Statistical uncertainty arises from the finite attack-set size, so additional independent attack traces can improve precision for the same family. The attacker family determines the scope of the guarantee. A precise bound for a weak family can therefore exclude stronger attackers [5,30]. Expanding the family addresses this limitation, but changes the recovery probability being bounded and may increase the certified value.

3.3 Three Certification Modes

BI supports three choices of attacker family. The singleton mode covers one model selected using validation traces, the finite-suite mode covers all members of a fixed trained suite, and the local-family mode extends coverage to models whose output scores remain close to those of the trained models. Each mode specifies the attacks covered by its confidence guarantee.

Singleton certificate $\mathrm{BI}^{\mathrm{sing}}$. Let $\widehat{h}_{\mathrm{sel}}$ be a model selected before the attack traces are used. Its success count on independent attack traces follows a binomial distribution, which we use to compute a one-sided upper confidence bound on its population recovery probability. We use the KL-binomial bound by default. Appendix B.3 gives its construction together with the Clopper–Pearson alternative and the simpler Hoeffding bound. Applying Definition 2 expresses the resulting probability bound in bits.

Finite-suite certificate $\mathrm{BI}^{\mathrm{suite}}$. Let $\mathcal{H}_{\mathrm{tr}} = \{h_1, \ldots, h_M\}$ contain all trained attacks covered by the evaluation. Different preprocessing choices, hyperparameters, and random seeds can produce different attacks, so the suite must include each resulting model whose success is to be covered. The entire suite is fixed independently of the attack set $S_{\mathrm{te}} = \{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{te}}}$.

For each model h_j, we compute the success count S_j and empirical recovery probability $\widehat{p}_j$ as

$$S_j := \sum_{i=1}^{n_{\mathrm{te}}} \mathbf{1}\{h_j(Y_i) = X_i\}, \qquad \widehat{p}_j := \frac{S_j}{n_{\mathrm{te}}}. \tag{13}$$

The quantity to certify is $P_{\mathrm{suite}} := \max_j \Pr[h_j(Y) = X]$, the maximum population recovery probability over the suite. The model that attains this maximum need not have the highest empirical recovery probability on an attack set. Thus, the confidence guarantee must cover every suite member simultaneously.

Inverting a Bernoulli concentration bound gives the upper and lower confidence endpoints for each attacker as

$$U_j := \sup\{q \in [\widehat{p}_j, 1] : n_{\mathrm{te}} D_{\mathrm{Ber}}(\widehat{p}_j \| q) \le \ln(2M/\delta)\}, \tag{14}$$

$$L_j := \inf\{q \in [0, \widehat{p}_j] : n_{\mathrm{te}} D_{\mathrm{Ber}}(\widehat{p}_j \| q) \le \ln(2M/\delta)\}, \tag{15}$$

where

$$D_{\mathrm{Ber}}(u\|v) := u \ln \frac{u}{v} + (1-u) \ln \frac{1-u}{1-v} \tag{16}$$

is the binary Kullback–Leibler divergence, with the usual limiting conventions at zero and one. For a fixed empirical recovery probability, this divergence determines the population recovery probabilities included in the confidence interval.

The factor $2M$ accounts for the upper and lower endpoints of all M models. We allocate failure probability $\delta/(2M)$ to each endpoint and apply a union bound to obtain simultaneous coverage with probability at least $1-\delta$. Consequently, taking the largest lower endpoint and the largest upper endpoint yields

$$p^{-}_{\mathrm{suite}} \le P_{\mathrm{suite}} \le p^{+}_{\mathrm{suite}}, \qquad p^{-}_{\mathrm{suite}} := \max_j L_j, \quad p^{+}_{\mathrm{suite}} := \max_j U_j. \tag{17}$$

This argument does not require the models' predictions on the same trace to be independent. Proposition 2 states the guarantee, with its proof in Appendix B.4.

We report the empirical maximum $p_{\mathrm{obs}} := \max_j \widehat{p}_j$ alongside these endpoints and express all three quantities in bits as

$$\begin{aligned} \mathrm{BI}^{-}_{\mathrm{suite}} &:= \log_2\bigl(\max\{1, K p^{-}_{\mathrm{suite}}\}\bigr)\,, \\ \mathrm{BI}^{\mathrm{obs}}_{\mathrm{suite}} &:= \log_2\bigl(\max\{1, K p_{\mathrm{obs}}\}\bigr)\,, \\ \mathrm{BI}^{+}_{\mathrm{suite}} &:= \log_2\bigl(\max\{1, K p^{+}_{\mathrm{suite}}\}\bigr)\,. \end{aligned} \tag{18}$$

The upper endpoint is the BI certificate. The empirical value describes performance on the attack set. On the same confidence event, a positive lower endpoint establishes that at least one suite member has population recovery probability above the random-guessing baseline. Reporting all three values distinguishes empirical performance from statistical uncertainty.

Local-family certificate $\mathrm{BI}^{\mathrm{loc}}(\varepsilon)$. The third mode extends the guarantee beyond the trained suite by constraining deviations in model output scores relative to trained reference models. The output scores are normalized on a fixed scale. A score radius ε then bounds the deviation in each class score from the corresponding reference score uniformly over the trace domain. The union of these neighborhoods forms the local family $\mathcal{H}_{\mathrm{loc}}(\varepsilon)$, defined formally in Section 4.2.

This restriction limits which classification errors can be corrected within a reference model's neighborhood. The classification margin is the true-class score minus the highest competing class score. If each class score changes by at most ε, the margin changes by at most 2ε. A trace with reference margin below -2ε therefore remains incorrectly classified throughout the neighborhood. The certificate counts traces whose reference margin is at least -2ε and applies a binomial confidence bound to this relaxed-margin count.

A larger radius covers a wider family and can increase the count, producing a larger upper bound on its recovery success. The resulting bound need not be attained by a model in the larger family. Section 4.2 gives the formal result and the confidence allocation for evaluating a radius grid fixed before certification.

Table 3: The three BI modes differ in the attackers covered by the guarantee.

Mode	Attacker family	Certification method
$\mathrm{BI}^{\mathrm{sing}}$	The family contains one model selected using validation traces.	A one-sided binomial bound certifies its recovery probability.
$\mathrm{BI}^{\mathrm{suite}}$	The family contains all models in a fixed trained suite.	The maxima of simultaneous binomial confidence endpoints bound the suite's maximum recovery probability.
$\mathrm{BI}^{\mathrm{loc}}(\varepsilon)$	The family contains models within a specified score radius of a trained reference model.	Binomial bounds on relaxed-margin counts provide simultaneous coverage over the reference models and the chosen radius grid.

At radius zero, models in each neighborhood have the same score function as their reference model and therefore make the same predictions under the same tie-breaking rule. The relaxed count also equals the ordinary correctness count when there are no ties for the largest score. When these counts are equal, the local and suite upper bounds agree if they use the same attack traces and allocate the same failure probability to each upper bound. Differences in tie handling or confidence allocation can therefore produce different upper bounds, as discussed in Remark 3.

The finite-suite mode is the main certificate used in our experiments, while the local extension examines how the bound changes when the evaluated family includes models beyond the trained suite. Their practical construction is given in Appendix B.1. In each case, the probability bound and the family it covers should be reported together.

3.4 Interpreting Tightness and Making a Security Decision

A certified upper bound can be large because the attacks succeed frequently or because the available attack set leaves substantial statistical uncertainty. To distinguish these cases, we report the empirical maximum together with both confidence endpoints and the statistical margin above the empirical maximum. Table 4 lists these quantities on the probability and BI scales.

When the empirical recovery probability exceeds the random-guessing baseline, the margin ratio is

$$R_{\mathrm{margin}} := \frac{p^{+}_{\mathrm{suite}} - p_{\mathrm{obs}}}{p_{\mathrm{obs}} - 1/K}. \tag{19}$$

A value below one indicates that the empirical recovery advantage exceeds the statistical margin, whereas a value above one indicates that the margin is larger. The ratio therefore measures the relative contribution of statistical uncertainty on the probability scale. When empirical success is at or below the random-guessing baseline, we report the confidence interval and its margin without interpreting the ratio because the denominator is nonpositive.

Table 4: The diagnostics quantify statistical uncertainty in the finite-suite certificate.

Quantity	Interpretation
$p^-_{\text{suite}},\ p_{\text{obs}},\ p^+_{\text{suite}}$	These are the lower confidence endpoint, empirical maximum, and upper confidence endpoint for recovery probability.
$p^+_{\text{suite}} - p_{\text{obs}}$	This difference is the statistical margin above the empirical maximum.
$\text{BI}^-_{\text{suite}},\ \text{BI}^{\text{obs}}_{\text{suite}},\ \text{BI}^+_{\text{suite}}$	These are the corresponding lower endpoint, empirical value, and upper endpoint in bits.
Δ_{BI}	$\Delta_{\text{BI}} = \text{BI}^+_{\text{suite}} - \text{BI}^{\text{obs}}_{\text{suite}}$ is the statistical margin in bits.
R_{margin}	R_{margin} is the ratio of the statistical margin to the empirical recovery advantage, defined below when $p_{\text{obs}} > 1/K$.

Three-way decision rule. The tightness report describes statistical uncertainty, while a security decision requires a maximum acceptable recovery probability p_{thr} chosen for the evaluation. On the simultaneous confidence event, the suite interval supports the following three outcomes.

- The result is **certified secure against the suite** if $p^+_{\text{suite}} \le p_{\text{thr}}$ because every attacker in the suite has recovery probability at most the threshold.
- The result is **certified vulnerable to a suite member** if $p^-_{\text{suite}} > p_{\text{thr}}$ because at least one attacker in the suite has recovery probability above the threshold.
- The result is **inconclusive** if $p^-_{\text{suite}} \le p_{\text{thr}} < p^+_{\text{suite}}$ because the interval includes values on both sides of the threshold.

The decision depends on the position of the confidence interval relative to p_{thr}. A large statistical margin relative to the empirical advantage therefore does not automatically imply an inconclusive result. Similarly, an upper bound of one provides no useful restriction on recovery probability, but a lower bound above p_{thr} still establishes vulnerability. These decisions apply to the fixed suite and do not extend to attackers outside it.

3.5 Ideal BI and Bayes Vulnerability

The preceding certificates bound recovery within a specified attacker family. For the same bound to cover unrestricted recovery, the family's supremum recovery probability must equal the Bayes-optimal recovery probability. This additional property is called *Bayes-completeness*.

A Bayes-optimal attacker applies the maximum a posteriori decision rule to the true posterior $\Pr[X = x \mid Y = y]$ and predicts the most probable secret value for each trace. Its success probability, i.e., unrestricted Bayes vulnerability, is

$$P_{\text{guess}}(X \mid Y) := \mathbb{E}_Y \max_{x \in \mathcal{X}} \Pr[X = x \mid Y]. \tag{20}$$

A family $\mathcal{H}$ is Bayes-complete for the distribution when

$$P^{\mathcal{H}}_{\text{guess}}(X \mid Y) = P_{\text{guess}}(X \mid Y). \tag{21}$$

This equality concerns the supremum recovery probability over the family and does not require a particular trained model to attain it. This condition therefore extends the certificate to unrestricted guessing success. When this condition is not established, the guarantee applies to the specified family.

The same condition permits an interpretation in terms of mutual information. A bound on Bayes-optimal recovery probability constrains the expected maximum posterior probability. This constraint gives a lower bound on conditional Shannon entropy and hence an upper bound on mutual information.

Lemma 1 (MI bound under Bayes-completeness). *Let X be uniform on an alphabet of size K, and suppose $\mathcal{H}$ is Bayes-complete. Let $\varepsilon^\star$ satisfy Equation (4), with $\overline{\varepsilon}^\star$ defined in Definition 2. On the same event of probability at least $1-\delta$,*

$$I(X;Y) \le \mathrm{BI}^{\mathcal{H}}_{\text{pop}}(X;Y) \le \beta(\overline{\varepsilon}^\star, K). \tag{22}$$

Conditional Shannon entropy is at least $-\log_2 P_{\text{guess}}(X \mid Y)$. It follows that

$$I(X;Y) \le \log_2(K P_{\text{guess}}(X \mid Y)). \tag{23}$$

Bayes-completeness allows us to replace unrestricted guessing success with the family-wide quantity and then apply the certificate. The conversion is deterministic and therefore preserves the original confidence level. Appendix C.4 gives the full derivation using the entropy and guessing-probability relations [28,14,10]. BI thus provides a recovery guarantee for the specified attacker family and, when Bayes-completeness is established, also bounds mutual information at the same confidence level.

3.6 A General Form for Family-Wide Certification

The singleton and suite constructions follow a general principle of uniform convergence. A bound on the generalization gap that holds simultaneously for every attacker in a family yields a certificate for the family's supremum recovery probability. The next proposition formalizes the principle introduced in Section 2.3.

Proposition 1 (Uniform confidence bound for an attacker family). *Let $\mathcal{H}$ be fixed independently of an attack set S_{te} containing $n_{\text{te}} \ge 1$ i.i.d. trace-label pairs. Write $\widehat{\Pr}_{S_{\text{te}}}[h(Y) = X]$ for the fraction of these traces that h classifies correctly. Suppose*

$$\sup_{h \in \mathcal{H}} \left| \Pr[h(Y) = X] - \widehat{\Pr}_{S_{\text{te}}}[h(Y) = X] \right| \le \Gamma(\mathcal{H}, n_{\text{te}}, \delta) \tag{24}$$

holds with probability at least $1-\delta$, where Γ may depend on S_{te}. On this event,

$$\mathrm{Adv}^{\mathcal{H}}(X \mid Y) \le \sup_{h \in \mathcal{H}} \widehat{\mathrm{Adv}}_{S_{\text{te}}}(h) + \Gamma(\mathcal{H}, n_{\text{te}}, \delta), \tag{25}$$

where $\widehat{\mathrm{Adv}}_{S_{\text{te}}}(h) := \widehat{\Pr}_{S_{\text{te}}}[h(Y) = X] - 1/K$.

Proof. On the stated event, each attacker's population recovery probability is at most its empirical recovery probability plus Γ. Subtracting $1/K$ and taking the supremum over $h \in \mathcal{H}$ gives Equation (25).

The proposition expresses the guarantee as empirical recovery advantage plus a statistical margin, following the uniform-convergence framework [39,31]. For a finite suite, simultaneous binomial confidence bounds provide this guarantee. Retaining the individual bounds yields the direct construction in Equation (14).

The local-family certificate in the next section extends coverage beyond the trained suite by bounding each model's recovery probability through the relaxed-margin event of a fixed reference model. Independent attack traces then provide a confidence bound on the probability of this event. This construction certifies the entire local family using only evaluations of its reference models. Increasing the attack-set size improves statistical precision, while the score radius controls the extent of family coverage.

4 Certification for Neural-Network Attacks

The finite-suite certificate bounds the maximum population recovery probability over a fixed suite of trained models. This section extends coverage to models whose output scores remain within a specified distance of a trained reference model. We first describe temporal locality as an architectural constraint and then define local score families. These families admit a margin-based certificate that uses only the reference models' scores on independent attack traces. We also explain how architecture-specific bounds establish score proximity and how recovery certificates can extend to multiple targets or traces.

The statistical certificate applies to multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and Transformers whose score functions satisfy a uniform score-distance condition. Architecture-specific perturbation bounds provide sufficient conditions for membership in this family by accounting for each model's layers and aggregation operations. For CNNs with shared filters and normalized average pooling, these bounds can avoid explicit dependence on trace dimension. This property depends on the input domain and parameter constraints, while MLPs and Transformers require bounds that account for their respective architectures.

4.1 Temporal Locality and the Trace Model

An attacker can process a high-dimensional trace by extracting features from local windows and aggregating them to predict the secret variable. We denote the temporal window length by τ, which specifies the number of consecutive sampling positions available to each local feature computation. Table 5 summarizes this structure for several side-channel attack strategies. Their aggregation rules determine which architecture-specific perturbation bounds apply.

We represent each trace as $Y \in \mathbb{R}^{T \times C_0}$, where T is the temporal dimension and C_0 is the number of measurement channels. The trace dimension is therefore

Table 5: Profiled side-channel attacks use local features with different aggregation rules. Membership in the family defined below additionally requires the specified additive aggregation.

Attack strategy	**Local input**	**Aggregation**
Template attack [9]	The input contains selected points of interest.	Recovery uses a model likelihood or combined log-likelihoods.
Sliding-window attack [34,5]	Each input is a window of trace samples.	Window outputs are combined by taking their maximum or average.
CNN-based attack [5,43,34]	Each local feature depends on the network's receptive field.	Pooling and an output layer combine the features.
Horizontal profiling [4]	Each window covers a repeated operation.	The attack combines evidence across repetitions.

$d = TC_0$, consistent with Section 2.1. For a single power or electromagnetic measurement channel, $C_0 = 1$, so the trace dimension equals its temporal dimension. The secret variable X is uniform over an alphabet $\mathcal{X}$ of size K, with $K = 256$ for an 8-bit intermediate.

The score-distance analysis requires fixed preprocessing that maps every admissible trace into a bounded domain $\mathcal{Y}_B$. Throughout this domain, the vectorized trace must satisfy $\|Y\|_2 \le B_Y$. This uniform norm bound must follow from a justified input-range constraint or an enforced preprocessing transformation. Centering and scaling can be included in preprocessing, but empirical norms from a finite dataset alone do not establish the required bound.

For a window length $\tau \in \{1, \dots, T\}$, we define the patch operator and patch count as

$$P_t(Y) := Y_{t:t+\tau-1,:}, \qquad N_\tau := T - \tau + 1. \tag{26}$$

The patch $P_t(Y)$ spans τ consecutive sampling positions across all measurement channels, beginning at position t. Its dimension is τC_0, and N_τ is the number of patches extracted from a trace. The following attacker family applies a shared feature map to these patches and averages the resulting features before predicting the secret variable.

Definition 3 (τ-local attacker family). *For a fixed window length τ, let $\mathcal{H}_\tau$ contain classifiers $h(Y) = \arg\max_{x \in \mathcal{X}} s_x(Y)$ with scores*

$$s_x(Y) = A\left(b_x + \frac{1}{N_\tau} \sum_{t=1}^{N_\tau} \phi_x(P_t(Y))\right), \qquad x \in \mathcal{X}. \tag{27}$$

Here ϕ_x is a shared patch feature map, b_x is a class-specific bias, and A is a fixed activation. The allowable feature maps and parameter constraints are specified as part of the family. Each model's parameters are fixed independently of the

attack traces used for certification, and a common deterministic rule resolves ties between class scores.

This definition specifies both local feature extraction and aggregation across the trace. A CNN with shared filters, the required receptive field, and normalized average pooling can implement this structure. Patch-based multilayer perceptrons (MLPs) and Transformers can also satisfy the definition when they use the same additive aggregation. Membership depends on the complete score function, so restricting only the first layer or attention window is insufficient. Models with other aggregation rules require a separate architecture analysis, although they can still be certified as fixed members of a finite suite.

The window length determines which trace coordinates a feature map can combine. For example, a product feature can combine several leakage samples within one patch, as in higher-order side-channel analysis. The leakage distribution determines whether these local features suffice for Bayes-optimal recovery. Temporal locality alone therefore implies neither a minimum window length for successful recovery nor an increase in certified recovery probability with window length. Appendix C.4 states the additional assumptions under which an expressive additive local family can match the Bayes-optimal recovery probability.

Temporal locality constrains the architecture through the family $\mathcal{H}_\tau$. The local score family introduced next instead constrains output deviations relative to reference models. Its margin certificate applies to any architecture satisfying the score-distance condition, without requiring membership in $\mathcal{H}_\tau$.

4.2 Local Score Families and the Margin Certificate

We first state the finite-suite guarantee described in Section 3.3. The trained suite supplies the reference models for the local-family extension. The suite may contain different architectures, provided each model and its preprocessing are fixed independently of the attack traces.

Proposition 2 (Finite-suite certificate). *Let $\mathcal{H}_{\mathrm{tr}} = \{h_1, \ldots, h_M\}$ be fixed independently of an attack set S_{te} of $n_{\mathrm{te}} \geq 1$ i.i.d. trace-label pairs, with $M \geq 1$. Let S_j count the correct predictions of h_j, and let $L_{\mathrm{binom}}(S, n, \alpha)$ and $U_{\mathrm{binom}}(S, n, \alpha)$ be the one-sided $(1-\alpha)$ KL-binomial endpoints of Appendix B.3. With probability at least $1 - \delta$, the following bounds hold simultaneously for every j.*

$$L_{\mathrm{binom}}\left(S_j, n_{\mathrm{te}}, \frac{\delta}{2M}\right) \leq \Pr[h_j(Y) = X] \leq U_{\mathrm{binom}}\left(S_j, n_{\mathrm{te}}, \frac{\delta}{2M}\right). \tag{28}$$

Taking the maximum of the lower and upper endpoints gives $p^-_{\mathrm{suite}} \leq P_{\mathrm{suite}} \leq p^+_{\mathrm{suite}}$, with the corresponding BI values defined in Equation (18).

The allocation $\delta/(2M)$ covers two one-sided events for each model. If only an upper bound is needed, the evaluator can instead use δ/M. The proof in Appendix B.4 applies a union bound to the individual confidence statements, so it does not require different models' predictions on a trace to be independent.

From trained models to a score neighborhood. A neural network assigns a class score to each candidate secret value and predicts the class with the maximum score. Since different parameter values can produce similar class scores, we define a neighborhood by bounding deviations in the score function. A fixed normalization specifies the scale on which these deviations are measured.

Definition 4 (Score normalization). *Each model h has a fixed score representation $s_h(Y) \in \mathbb{R}^K$ and predicts $h(Y) = \arg\max_x s_{h,x}(Y)$ using a common deterministic tie rule. For each reference model h_j, choose a positive constant A_j independently of the attack traces. When comparing a model with h_j, define its normalized score as*

$$\bar{s}_h^{(j)}(Y) := \frac{s_h(Y)}{A_j}. \tag{29}$$

The same constant is used for the candidate and reference scores in this comparison. All margins and radii for neighborhood j are measured on this scale.

Remark 2 (Choosing the score scale). Multiplying all class scores by the same positive constant leaves the prediction unchanged but changes numerical score differences. The score representation and its normalization must therefore be fixed before evaluation. A normalization selected using profiling or validation traces defines a score scale but does not automatically establish a uniform bound on score magnitude. Consequently, comparisons across architectures must specify the normalization conventions because equal numerical radii can represent different changes in model behavior.

Definition 5 (Local score family). *Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_M)$ contain nonnegative score radii for the fixed reference models. Define*

$$\begin{aligned} \mathcal{H}_{\mathrm{loc}}(\varepsilon) &:= \bigcup_{j=1}^{M} \mathcal{H}_j^{\mathrm{loc}}(\varepsilon_j), \\ \mathcal{H}_j^{\mathrm{loc}}(\varepsilon_j) &:= \left\{ h : \sup_{Y \in \mathcal{Y}_B} \left\| \bar{s}_h^{(j)}(Y) - \bar{s}_{h_j}^{(j)}(Y) \right\|_\infty \le \varepsilon_j \right\}. \end{aligned} \tag{30}$$

The infinity norm takes the largest absolute difference across class scores. Thus, a model belongs to neighborhood j only if every normalized class score differs from the reference score by at most ε_j on every admissible trace.

Increasing a radius expands the corresponding score neighborhood. At radius zero, every member has the same score function as its reference model on $\mathcal{Y}_B$. These score radii differ from $\varepsilon^\star$, which bounds the population recovery advantage. Empirical score differences can inform radius selection, but establishing family membership requires a uniform bound over the admissible trace domain.

Why a score radius leads to a success bound. A model in a score neighborhood can correct a reference error only when the permitted perturbation allows the true class to attain the maximum score. If each class score changes by at most ε_j,

the difference between the true-class score and any competing score changes by at most $2\varepsilon_j$. We therefore count every attack trace on which the reference model's true-class score is no more than $2\varepsilon_j$ below the highest competing score. This relaxed-margin count includes all correct reference predictions and every trace that a model in the neighborhood could classify correctly. It can also include traces that no such model recovers, so the resulting bound may be conservative.

Theorem 1 (Local-family margin certificate). *Let $S_{\mathrm{te}} = \{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{te}}}$ contain $n_{\mathrm{te}} \geq 1$ i.i.d. attack pairs, with traces supported on $\mathcal{Y}_B$. Fix the suite, preprocessing, score representations, positive normalization constants, and a finite nonempty grid $\mathcal{E}$ of radius vectors independently of S_{te}. For reference model h_j, define the normalized margin*

$$\rho_j(x, y) := \bar{s}^{(j)}_{h_j,x}(y) - \max_{a \neq x} \bar{s}^{(j)}_{h_j,a}(y). \tag{31}$$

For each $\varepsilon \in \mathcal{E}$, define the relaxed-margin count

$$S_j^{\mathrm{loc}}(\varepsilon_j) := \sum_{i=1}^{n_{\mathrm{te}}} \mathbf{1}\{\rho_j(X_i, Y_i) \geq -2\varepsilon_j\}. \tag{32}$$

With probability at least $1-\delta$, the following bound holds simultaneously for every $\varepsilon \in \mathcal{E}$.

$$\begin{aligned} \sup_{h \in \mathcal{H}_{\mathrm{loc}}(\varepsilon)} \Pr[h(Y) = X] &\leq p^+_{\mathrm{loc}}(\varepsilon), \\ p^+_{\mathrm{loc}}(\varepsilon) &:= \max_{j \leq M} U_{\mathrm{binom}}\left(S_j^{\mathrm{loc}}(\varepsilon_j), n_{\mathrm{te}}, \frac{\delta}{M|\mathcal{E}|}\right). \end{aligned} \tag{33}$$

The proof first establishes that a correct prediction by any model in neighborhood j implies a reference margin of at least $-2\varepsilon_j$. The probability of this relaxed-margin event therefore bounds the recovery probability of every model in that neighborhood. A binomial upper confidence bound then bounds this event probability using its empirical count. Allocating failure probability across the M reference models and $|\mathcal{E}|$ radius vectors then gives simultaneous coverage by a union bound. This argument requires no independence across models or grid values, as shown in Appendix C.2.

The simultaneous guarantee allows the evaluator to report any radius vector from the fixed grid. A separate calibration partition may be used to select a representative radius, provided the final attack traces used for certification remain independent of that selection. The protocol and confidence allocation must distinguish this calibration data from the traces used to compute the final certificate.

Remark 3 (Recovering the suite certificate). At radius zero, the relaxed event is $\rho_j(X_i, Y_i) \geq 0$. This includes ties in which the true class has a maximal score even if the deterministic tie rule chooses a different class. Consequently, absence of ties for the largest score is a sufficient condition for the relaxed count to equal

Table 6: Architecture-specific perturbation bounds establish local-family membership when they hold uniformly over the admissible input domain.

Architecture	Sufficient score-distance analysis
MLP	The bound accounts for weight and bias perturbations in each layer and their propagation through subsequent layers.
CNN	The bound accounts for perturbations in the shared patch network and output layer under normalized pooling. Its dependence on trace dimension is determined by the patch-domain and parameter bounds.
Transformer	The bound accounts for block perturbations and their subsequent propagation. The analysis includes tokenization, attention, residual connections, normalization, and pooling.

the ordinary correctness count. When the counts agree, equality of the suite and local upper endpoints additionally requires the same attack traces and the same per-endpoint confidence allocation. The usual allocations $\delta/(2M)$ for the two-sided suite interval and $\delta/(M|\mathcal{E}|)$ for the local grid need not coincide.

Definition 6 (Local-family BI certificate). *Under Theorem 1, let*

$$\varepsilon^{\star}_{\mathrm{loc}}(\varepsilon) := [p^{+}_{\mathrm{loc}}(\varepsilon) - 1/K]_{+}, \qquad [a]_{+} := \max\{a, 0\}. \tag{34}$$

The local-family BI certificate is

$$\mathrm{BI}^{\mathrm{loc}}(\varepsilon) := \min\left\{\log_2(1 + K\varepsilon^{\star}_{\mathrm{loc}}(\varepsilon)), \log_2 K\right\}. \tag{35}$$

It bounds the population BI of $\mathcal{H}_{\mathrm{loc}}(\varepsilon)$ on the confidence event of the theorem, using the reporting convention in Definition 2.

4.3 Establishing the Score Radius

The margin theorem certifies a family defined by a uniform score-distance condition. To cover a specified set of parameter perturbations, we must bound their effect on the output scores throughout $\mathcal{Y}_B$. A deterministic bound Δ_j on the unnormalized score difference then gives the normalized radius $\varepsilon_j = \Delta_j/A_j$. Table 6 summarizes sufficient approaches for three architectures, with further details in Appendix C.3.

MLP. For an MLP, a parameter perturbation changes the output of one layer and propagates through subsequent layers. The operator norm of each weight matrix bounds the amplification caused by its linear transformation. Accounting for the activations and summing the propagated contributions therefore bounds the final score difference.

Consider an L-layer reference model with zero biases and one-Lipschitz activations satisfying $\sigma(0) = 0$. This condition means that a zero input produces a zero output, so the activation introduces no constant offset. Together with

the one-Lipschitz property, it ensures that $\|\sigma(z)\|_2 \le \|z\|_2$, allowing intermediate activation norms to be bounded using the input norm and weight operator norms. Common activations such as ReLU and tanh satisfy these conditions. Let $B_{j,\ell}$ bound the reference weight operator norm and let r_ℓ bound the weight perturbation in layer ℓ. Under these assumptions, a sufficient bound is

$$\sup_{Y\in\mathcal{Y}_B} \|s_h(Y)-s_{h_j}(Y)\|_\infty \le B_Y \sum_{\ell=1}^{L} r_\ell \left(\prod_{q<\ell}(B_{j,q}+r_q)\right)\left(\prod_{q>\ell} B_{j,q}\right) =: \Delta_j^{\mathrm{MLP}}(r). \tag{36}$$

The first product bounds the input reaching the perturbed layer, and the second bounds propagation through the remaining reference layers. Nonzero biases contribute to the intermediate activation bounds through the recursion in Lemma 2. Activations that do not map zero to zero additionally require their offsets to be included. Dividing the resulting score bound by A_j gives the normalized radius.

CNN. For a CNN with shared patch processing and normalized average pooling, the unnormalized score vector has the form

$$s_W(Y) = W_{\mathrm{out}}\left(\frac{1}{N_\tau}\sum_{t=1}^{N_\tau}\phi_W(P_t(Y))\right) + b_{\mathrm{out}}. \tag{37}$$

The MLP argument can be applied to the shared patch network using the patch bound $B_{\mathrm{patch}} := \sup_{Y\in\mathcal{Y}_B,t}\|P_t(Y)\|_2$. If the feature difference is at most Δ_{patch} on every patch, normalized averaging preserves this bound because

$$\left\|\frac{1}{N_\tau}\sum_t\big(\phi_W(P_t(Y)) - \phi_{W_j}(P_t(Y))\big)\right\|_2 \le \frac{1}{N_\tau}\sum_t \Delta_{\mathrm{patch}} = \Delta_{\mathrm{patch}}. \tag{38}$$

Normalized averaging therefore eliminates the patch count as an additional multiplicative factor. The complete score bound must also account for the output layer and its parameter perturbations. Dependence on trace dimension can still enter through the patch-domain or parameter bounds. Flattening and unnormalized aggregation require separate bounds because the normalized-average argument does not apply to these operations, as discussed in Lemma 3.

Transformer. For a Transformer, the perturbation analysis proceeds at the block level and includes attention, residual connections, normalization, and the output head. The tokenization, sequence length, and attention pattern must be fixed, and every intermediate activation arising in the perturbation argument must lie in a domain on which the stated bounds hold.

For a Transformer with B blocks, let $\Delta_{j,b}(r_b)$ bound the change produced by perturbing block b on the relevant activation domain. Let $L_{j,q}$ bound amplification through each later reference block, let $L_{j,\mathrm{out}}$ bound amplification through the complete reference output head, and let $\Delta_{j,\mathrm{head}}$ bound the direct effect of

perturbing that head. The corresponding telescoping bound takes the form

$$\sup_{Y\in\mathcal{Y}_B} \|s_h(Y)-s_{h_j}(Y)\|_\infty \le \Delta_{j,\mathrm{head}}+L_{j,\mathrm{out}}\sum_{b=1}^{B}\Delta_{j,b}(r_b)\prod_{q>b}L_{j,q} =: \Delta_j^{\mathrm{Tr}}(r). \quad (39)$$

Each block difference is measured on inputs produced by the preceding perturbed blocks, and the head difference is bounded on the final perturbed activation domain. The explicit output-head factor distinguishes amplification of upstream changes from changes to the head's own parameters. Lemma 4 describes this blockwise analysis. Its constants must account for all these contributions to establish the normalized radius $\Delta_j^{\mathrm{Tr}}(r)/A_j$.

Interpreting the local certificate. The construction supports two uses. An evaluator can specify a score radius and certify the family defined by the corresponding uniform constraint. Alternatively, the evaluator can specify parameter perturbations and derive a sufficient radius using a deterministic bound. Empirical score differences can guide the first choice, but a uniform perturbation bound is required to establish that perturbed models belong to the certified family.

The resulting BI curve shows how the certified upper bound varies with the score radius. Incorrect reference predictions with small negative margins enter the relaxed count even at a small radius. A low empirical recovery probability for the reference model can therefore coexist with a large local-family bound. If the relaxed count equals the attack-set size for any reference model, its upper endpoint is one, which makes the certificate for the union vacuous. The radius sweep must therefore be interpreted together with the reference models' empirical recovery probabilities and margin distributions, as illustrated in Section 5.6.

4.4 Composition across Targets and Traces

The certificates above concern recovery of one secret variable from one trace. A key-recovery attack may instead combine predictions for several target variables, accumulate evidence across several traces, or apply a search procedure to the combined scores. Extending a single-trace certificate to these settings requires either a valid composition or direct evaluation of the recovery procedure.

Let K denote the master key, T_i the i-th target variable with alphabet $\mathcal{T}_i$, and Z_i the probability vector output by its trained model. Write $Z = (Z_1, \ldots, Z_\nu)$ for the combined evidence from ν targets, with public inputs treated as fixed or included in the conditioning. The symbol K continues to denote the target alphabet size used earlier. A certificate for the top-scoring prediction alone need not bound what can be recovered from the full probability vector, because a downstream attack can use information in all its entries.

Post-processing of combined evidence. A soft-analytical side-channel attack (SASCA) combines model outputs with constraints from the cipher, often using belief propagation (BP) on a factor graph. Once the graph, schedule, public information, and algorithmic choices are fixed, its output is a deterministic function of Z.

Any key guess based on that output can also be made from Z by first running the same procedure, which gives

$$P_{\text{guess}}(\mathsf{K} \mid \text{BP}_{\mathsf{sched}}(Z)) \leq P_{\text{guess}}(\mathsf{K} \mid Z). \tag{40}$$

Lemma 5 proves this inequality without requiring independence among the evidence components. Thus, deterministic post-processing cannot increase unrestricted guessing probability. Applying this inequality to individual target certificates additionally requires a valid bound for the combined evidence.

Composition across targets. For divide-and-conquer composition, suppose the master key is represented by a tuple of independent uniform targets or an equivalent tuple under a known bijection. Assume also that the evidence vectors are independent and that the joint posterior factorizes into the per-target posteriors. If each per-target certificate bounds unrestricted guessing from its corresponding evidence vector by

$$P_{\text{guess}}(T_i \mid Z_i) \leq q_i, \qquad q_i := \min\{1, 1/|\mathcal{T}_i| + \varepsilon_i^\star\}, \tag{41}$$

then the following bounds hold on the joint confidence event.

$$P_{\text{guess}}(\mathsf{K} \mid Z) \leq \prod_{i=1}^{\nu} q_i, \qquad \widetilde{H}_\infty(\mathsf{K} \mid Z) \geq \sum_{i=1}^{\nu} -\log_2 q_i. \tag{42}$$

Here $\widetilde{H}_\infty$ is conditional min-entropy defined through guessing probability. The terms $-\log_2 q_i$ lower-bound the remaining conditional min-entropy. BI instead expresses the recovery bound relative to the random-guessing baseline, so it measures a different quantity. If the individual failure probabilities are δ_i, the joint confidence is at least $1 - \sum_i \delta_i$, as shown in Lemma 6.

Dependencies in a coupled SASCA can prevent this product decomposition. An alternative bound applies when the master key is uniform, each target is determined by the key and fixed public inputs, the evidence is conditionally independent through the targets, and valid per-target mutual information bounds are available. Under these assumptions, Lemma 7 combines the information bounds and applies Pinsker's inequality to bound key-recovery probability. Bayes-complete certificates can provide the required mutual information bounds, whereas arbitrary model classification accuracy alone is insufficient. If the composition assumptions are not established, the post-processing inequality remains valid but does not yield a per-target composition rule.

Composition across traces. When several traces correspond to the same secret, Proposition 5 gives an upper bound on recovery from their combined information, provided the traces are conditionally independent given the secret and the single-trace certificate supports the required MI bound. If the traces also have identical conditional distributions, the result uses the trace count and a common single-trace information bound. Otherwise, it uses the sum of the individual information bounds. The resulting curve bounds recovery probability under the stated assumptions and need not describe the empirical recovery rate of a particular attack.

Direct certification of a complete attack. If the evaluation concerns a fixed recovery procedure, its final outcome can be certified directly. One statistical trial is then a complete attack campaign containing the required traces and ending in a key-recovery success or failure. The models, preprocessing, combination of scores, search budget, and other pipeline choices must be fixed independently of the campaigns used for certification.

Proposition 3 (End-to-end pipeline certificate). *Let $\mathcal{A}_1, \ldots, \mathcal{A}_{n_c}$ be i.i.d. attack campaigns drawn from the evaluation distribution independently of all data used to train, select, or tune a fixed pipeline. Let S_i indicate whether that pipeline recovers the key in campaign $\mathcal{A}_i$. Then the indicators are i.i.d. Bernoulli variables with mean p_{pipe}. With probability at least $1-\delta$, the population recovery probability satisfies*

$$p_{\mathrm{pipe}} \leq U_{\mathrm{binom}}\left(\sum_{i=1}^{n_c} S_i, n_c, \delta\right), \tag{43}$$

where U_{binom} is the KL-binomial or Clopper–Pearson upper confidence bound.

The proof in Appendix C.8 treats the complete pipeline as a fixed attacker and each campaign as one Bernoulli trial. Direct certification therefore avoids intermediate composition bounds, with statistical precision determined by the number of independent campaigns. Its guarantee applies to the specified pipeline. Appendix C.9 reports a related empirical four-byte recovery diagnostic on raw ASCAD traces. The next section evaluates the finite-suite and local-family certificates on measured side-channel traces.

5 Experimental Evaluation

We evaluate the finite-suite and local-family BI certificates on eight side-channel benchmarks covering protected and unprotected AES implementations. These benchmarks include traces with desynchronization and random delays, as well as the CHES-CTF-2025 challenge. The experiments examine the statistical precision of the certificates, their dependence on trace dimension, and the effect of expanding the attacker family through score neighborhoods. We then relate the single-trace results to empirical key-recovery attacks that combine evidence from several traces. All the relevant code and a detailed user manual for replicating the experiments in this work are available at https://github.com/SEAL-IIT-KGP/PAC-Certification-of-Multivariate-Side-Channel-Traces.

5.1 Evaluation Protocol and Attacker Families

Each benchmark targets exact recovery of an 8-bit secret intermediate, so the secret alphabet has size $K = 256$ and random guessing succeeds with probability $1/256$. The six benchmarks comprising ASCAD with desynchronization levels 0, 50, and 100, AES-HD, AES-RD, and DPAv4 each use a TCHES20-style suite of 80 trained models. Each suite contains 40 MLP and 40 CNN architectures [43].

ASCAD random key and CHES-CTF-2025 each use a separate suite of 11 trained EstraNet models [18]. The selected-model experiments later in this section also evaluate Transformers, which are separate from the 80-model suites.

The certification protocol separates model development from evaluation. Candidate models are trained on profiling traces and compared on validation traces, and the evaluated suite is fixed before certification. For the six TCHES20 benchmarks, the attack traces are separate from the training and validation traces. AES-RD uses the first 12,500 attack traces, while the other five benchmarks use all available attack traces. Table 7 reports the trace dimension d, which is the number of values in each trace supplied to the model, and the attack-set size n_{te}, which is the number of traces used for evaluation. CHES-CTF-2025 is an exception to the required separation because its preprocessing uses the mean of the attack traces. We therefore report its computed lower and upper bounds as numerical diagnostics.

We use a failure probability of $\delta = 10^{-6}$ for each reported certification procedure. The finite-suite bounds account for all M models, and the local-family bounds additionally account for the fixed radius grid $\mathcal{E}$. This confidence allocation has no explicit dependence on trace dimension, although a change in trace dimension can affect the success counts and the resulting bounds. The confidence level of $1 - \delta$ applies separately to each procedure. A simultaneous guarantee across all benchmarks and experiments would require an additional allocation of the failure probability.

The main results use the finite-suite and local-family certificates from Proposition 2 and Theorem 1. Appendix D complements these results with singleton certificates for models selected using validation traces. Appendix D.5 reports synthetic experiments with a known Bayes rule and comparisons with additional attackers on measured traces. These comparisons assess the chosen suite against specific alternatives, but they do not establish Bayes-completeness.

5.2 Finite-Suite Results

Table 7 reports the maximum empirical recovery probability p_{obs} together with the lower and upper confidence bounds p^- and p^+. These bounds apply to the maximum population recovery probability among the fixed suite members. A BI upper bound of b bits therefore limits that recovery probability to at most $2^b/256$. Under this conversion, zero bits corresponds to the random-guessing baseline, while eight bits gives the trivial upper bound of one. All of these quantities concern recovery of the specified secret intermediate from one trace.

Across the six TCHES20 benchmarks, the BI upper bounds range from 1.124 bits on AES-HD to 5.351 bits on AES-RD. On AES-RD, the maximum empirical recovery probability is 13.968%, while its upper confidence bound is 15.946%. The difference reflects statistical uncertainty from the finite attack set. Thus, the certified upper bound includes a margin above the empirical recovery probability.

The ratio R_{margin} compares the statistical margin $p^+ - p_{\mathrm{obs}}$ with the empirical recovery advantage $p_{\mathrm{obs}} - 1/256$. It is below one for AES-RD and ASCAD random key, where the empirical advantage exceeds the statistical margin, and above

Table 7: The table reports finite-suite BI results for each trace dimension d. The suite superscript is omitted from the column headers. EPI is evaluated on the same attack traces using the model with the highest empirical recovery probability. Blue rows have $R_{\text{margin}} < 1$, while yellow rows have $R_{\text{margin}} > 1$. †The CHES-CTF-2025 row shows BI's numerical behavior when preprocessing uses the mean of the attack traces. Extending these results to a PAC guarantee requires establishing independence between preprocessing and the traces used for certification.

Dataset	d	M	n_{te}	p_{obs}	$[p^-, p^+]$	BI_{obs}	$[\text{BI}^-, \text{BI}^+]$	R_{margin}	EPI
ASCAD d0	700	80	10,000	0.00630	[0.00260, 0.01247]	0.690	[0.000, 1.675]	2.58	≈ 0
ASCAD d50	700	80	10,000	0.00970	[0.00485, 0.01701]	1.312	[0.311, 2.123]	1.26	0.239
ASCAD d100	700	80	10,000	0.00980	[0.00492, 0.01714]	1.327	[0.332, 2.134]	1.25	0.208
ASCAD random	1,400	11	100,000	0.01693	[0.01467, 0.01941]	2.116	[1.909, 2.313]	0.19	0.524
AES-HD	1,250	80	25,000	0.00520	[0.00288, 0.00851]	0.413	[0.000, 1.124]	2.56	-0.006
AES-RD	3,500	80	12,500	0.13968	[0.12135, 0.15946]	5.160	[4.957, 5.351]	0.15	3.097
DPAv4	4,000	80	500	0.02400	[0.00201, 0.09144]	2.619	[0.000, 4.549]	3.36	2.251
CHES-CTF-2025†	3,000	11	100,000	0.00429	[0.00320, 0.00561]	0.135	[0.000, 0.521]	3.43	-0.105

one for the other six rows. DPAv4 illustrates the uncertainty associated with a small attack set. Its maximum empirical recovery probability is 2.4% over 500 traces, but the upper confidence bound is 9.144%. Additional independent attack traces can improve statistical precision for the fixed suite. The scope of the guarantee still depends on which attackers the suite contains. Appendix D.4 further examines statistical margins and attacker-family size.

5.3 Comparison with Probability-Based Metrics

The EPI column reports a probability-based metric for the model with the highest empirical recovery probability in each suite, evaluated on the same attack traces. With the model weights fixed, we fit an inverse temperature over $[0, 64]$ using up to 10,000 profiling traces sampled without replacement. These calibration traces may overlap the original training data, but they are separate from the attack traces. EPI therefore provides a descriptive comparison for the selected model, whereas the BI confidence bounds account for all models in the suite. Since the two metrics measure different quantities, their values need not agree even though both are expressed in bits.

To examine the effect of probability calibration, we apply temperature scaling to the same fixed models at temperatures from 0.05 to 20 on six benchmarks. This transformation changes the predicted probabilities while preserving every predicted label. Raw PI varies by up to 27.31 bits, whereas recalibrated EPI varies by less than 10^{-9} bits. The empirical BI value remains unchanged because the exact-recovery outcomes are identical. The finite-suite confidence bounds also remain unchanged when the suite, attack-set size, and confidence allocation are fixed. This experiment shows how probability scaling can affect a metric without changing classification accuracy. Appendix D.7 provides the additional values for each benchmark.

Table 8: The table reports key-byte recovery success (%) across 100 shuffled orders of the attack traces used in Table 7. For each benchmark, the model with the highest empirical recovery probability is evaluated using the trace counts in the column headings. The results describe empirical performance on a fixed attack set. A dash indicates that the requested trace count exceeds the available attack-set size.

	Number of accumulated attack traces			
Benchmark	100	500	1,000	2,000
ASCAD d0	0	0	0	1
ASCAD d50	49	100	100	100
ASCAD d100	67	100	100	100
ASCAD random	98	100	100	100
AES-HD	1	4	13	30
AES-RD	100	100	100	100
DPAv4	100	100	—	—

5.4 Relation to Key-Recovery Attacks

We next examine how the same models perform when a key-recovery attack combines probability scores from several traces. For each of the six TCHES20 benchmarks and ASCAD random key, we select the model with the highest empirical recovery probability and reuse the attack traces from Table 7. This evaluates key-byte recovery on seven benchmarks, including the 80-model TCHES20 suites and the 11-model EstraNet suite for ASCAD random key. For each candidate key byte, the benchmark's evaluation procedure determines the corresponding intermediate label on each trace. We sum the model's log-probabilities for these labels across traces and rank the candidate key bytes by the resulting scores.

Table 8 reports the fraction of 100 randomly shuffled trace orders in which the correct key byte ranks first. Each order draws traces without replacement from the same fixed attack set, and ties in the accumulated scores are resolved by choosing the smaller key index. The repeated orders therefore measure empirical performance on the available traces. Because they reuse the same attack set, they do not constitute the independent campaigns required for the population certificate in Proposition 3.

At 500 traces, the ASCAD d50 and d100 models recover the correct key byte in all 100 orders, while the ASCAD d0 model succeeds in none. The ASCAD random-key model succeeds in 98 of the 100 orders at 100 traces and in every order at 500, 1,000, and 2,000 traces. AES-RD and DPAv4 achieve recovery in every order with only 100 traces. These comparisons relate the single-trace results to a concrete key-byte attack using the same models and attack traces. Key recovery also depends on the probabilities assigned to candidate intermediate values and on how this evidence accumulates across traces. The single-trace BI value alone therefore does not determine the number of traces required for key recovery.

Appendix C.9 extends the empirical evaluation to four-byte recovery on raw ASCAD traces. The attack combines fixed models for the individual bytes with candidate-key enumeration. It recovers all four bytes in 191/250 campaigns at 500 traces, 206/250 at 1,000 traces, and 221/250 at 2,000 traces, as reported in Table 10. This experiment uses different trace windows and models from the single-trace evaluation. It therefore provides additional evidence of key-recovery performance, but it does not directly relate the BI values in Table 7 to the number of traces required for four-byte recovery.

5.5 Dependence on Trace Dimension

The trace dimensions in Table 7 range from 700 to 4,000. Figure 1 examines a separate CHES-CTF-2025 experiment in which the trace dimension increases from 100 to 7,000. The computed BI upper bounds remain approximately between 0.3 and 0.4 bits over this range. In comparison, the GKOV mutual information estimates vary non-monotonically and become negative at the largest trace dimension, while the PI estimates decrease.

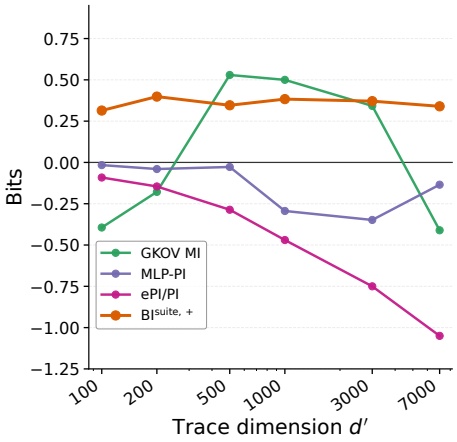


Fig. 1: The curves show metric values on CHES-CTF-2025 as the trace dimension increases from 100 to 7,000. The BI curve illustrates numerical stability when preprocessing uses the mean of the attack traces. Extending these results to a PAC guarantee requires establishing independence between preprocessing and the traces used for certification.

The CHES experiment illustrates that the computed BI upper bounds remain numerically stable over the evaluated trace dimensions. A PAC interpretation additionally requires preprocessing fixed independently of the certification traces. Extending the guarantee to unrestricted recovery also requires Bayes-completeness of the attacker family. The theoretical absence of explicit dependence on trace dimension concerns the binomial confidence calculation. Once the success counts, attack-set size, suite size, and confidence allocation are specified, this calculation requires no additional term involving trace dimension. Increasing the trace dimension can still affect the features learned by the models and their empirical recovery probabilities.

5.6 Local-Family Results

The local-family experiments examine how the upper bound changes when the certificate covers score neighborhoods around the trained reference models. Within each neighborhood, every normalized class score differs from its reference score

by at most ε on every admissible trace. Increasing ε therefore expands the attacker family, including models that were not trained or individually evaluated. Membership is defined by score differences over the admissible trace domain. To cover a specified parameter neighborhood, the evaluator must additionally establish a uniform bound on the score changes induced by those parameter perturbations, as described in Section 4. For the experiments below, a common radius is applied to all reference models in each evaluated family. The relaxed-margin counts then give the following probability and BI upper bounds through the binomial construction in Section 4.

$$p^{+}_{\mathrm{loc}}(\varepsilon) = \max_{j \le M} U_{\mathrm{binom}}\Big(S^{\mathrm{loc}}_{j}(\varepsilon),\, n_{\mathrm{te}},\, \tfrac{\delta}{M|\mathcal{E}|}\Big), \tag{44}$$

$$\mathrm{BI}^{\mathrm{loc},+}(\varepsilon) = \log_2\big(1 + K\,[\,p^{+}_{\mathrm{loc}}(\varepsilon) - 1/K\,]_{+}\big). \tag{45}$$

Here $S^{\mathrm{loc}}_{j}(\varepsilon)$ counts attack traces for which the reference model's normalized true-class score minus its highest competing score is at least -2ε. Every trace correctly classified by a model in the neighborhood must satisfy this condition. The count therefore provides an upper bound on the number of possible successes, although a model that achieves all of these successes need not exist. As the radius increases, the relaxed-margin counts and their upper bounds cannot decrease.

Radius selection and model normalization. The selected-model experiment uses one MLP, one CNN, and one Transformer on each of AES-RD, ASCAD d50, ASCAD d100, and ASCAD random key. For these twelve reference models, the reserved traces are divided into a calibration partition containing 20% of the traces and a certification partition containing the remaining 80%. The radius grid contains nine values from zero to 0.03 and is fixed before certification, as specified in Equation 103.

The reported radius $\varepsilon_{\mathrm{cal}} = 3 \cdot 10^{-4}$ is the smallest grid value satisfying two calibration criteria. First, the relaxed-margin count must increase relative to radius zero for at least 75% of the reference models. Second, the radius must be at least as large as the maximum normalized score difference observed when inference is repeated with different batch sizes. The first criterion holds at 10^{-4}, but the observed score difference of $1.905 \cdot 10^{-4}$ requires the next grid value, $3{\cdot}10^{-4}$. We then compute the certificate using the separate certification partition. Appendix D.6 describes the procedure in detail.

These calibration criteria select a radius based on the evaluated models and their implementation. To establish that another model belongs to the resulting family, its normalized score differences must satisfy the uniform condition in Definition 5. The score normalization also affects how the radius should be interpreted. The Transformer models use raw logits divided by a fixed constant determined from the training partition, whereas the MLP and CNN models use normalized probability scores. Consequently, equal numerical radii need not represent equivalent changes in model behavior across architectures.

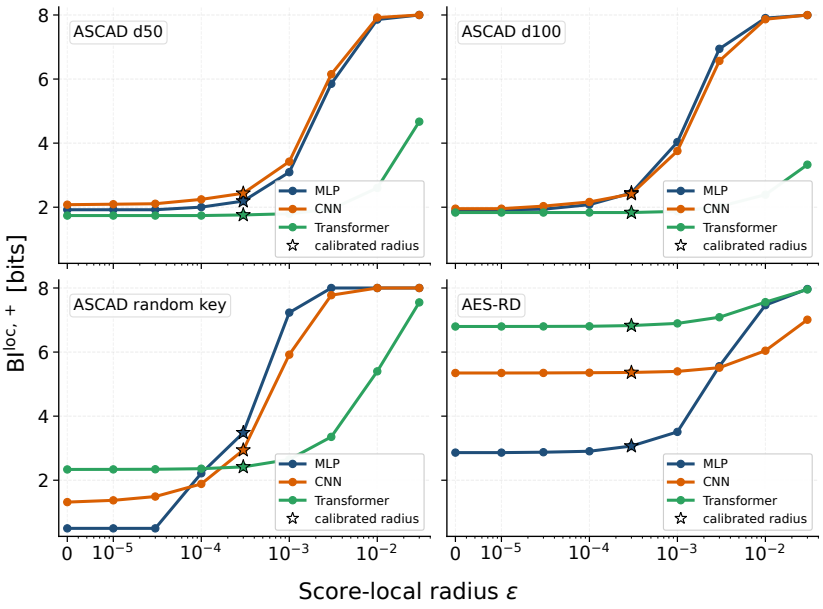


Fig. 2: The curves show local-family BI upper bounds for separate neighborhoods around twelve selected reference models. At $\varepsilon_{\mathrm{cal}} = 3 \cdot 10^{-4}$, all upper bounds are below eight bits. The ASCAD upper bounds range from 1.76 to 3.48 bits. On AES-RD, the upper bounds for the neighborhoods around the MLP, CNN, and Transformer are 3.07, 5.36, and 6.83 bits, respectively.

Interpreting the radius curves. Figure 2 shows that the upper bounds change little at small radii and then increase as more traces satisfy the relaxed-margin condition. A nearly constant bound indicates that expanding the score neighborhood adds few traces to the relaxed-margin count. When this count equals the number of certification traces, the BI upper bound reaches eight bits and gives the trivial upper bound of one on recovery probability. An increase at a smaller radius indicates that more traces have small negative true-class margins under the chosen normalization.

At the calibrated radius, the BI upper bounds for all twelve selected-model neighborhoods remain below eight bits. On AES-RD, the upper bounds are 3.07, 5.36, and 6.83 bits for the neighborhoods around the selected MLP, CNN, and Transformer, respectively. These results demonstrate that the margin certificate applies across all three architectures. Each value bounds a different model neighborhood, so their ordering alone does not rank the reference models by empirical recovery probability. These comparisons also depend on the score normalization and statistical uncertainty associated with each model.

The smaller grid radius $\varepsilon = 10^{-4}$ provides an additional comparison across the same twelve reference models. At this radius, all upper bounds remain below

Table 9: The table reports local-family upper bounds for the union of neighborhoods around 80 fixed MLP and CNN reference models at five of the nine radius values. AES-RD uses 12,500 certification traces, and each ASCAD benchmark uses 10,000. Each binomial upper bound uses failure probability $\delta/(80|\mathcal{E}|)$. The count S^{loc}, rate $\widehat{p}^{\,\mathrm{loc}}$, and architecture refer to a model whose neighborhood determines the maximum upper bound. Gray cells indicate that the relaxed-margin count equals the number of certification traces.

Benchmark	ε	S^{loc}	$\widehat{p}^{\,\mathrm{loc}}$	$\mathrm{BI}^{\mathrm{loc},+}$	Arch.
AES-RD	0	1,746	0.140	5.36	CNN
	10^{-4}	1,756	0.140	5.37	CNN
	$3 \cdot 10^{-4}$	1,767	0.141	5.37	CNN
	10^{-3}	1,812	0.145	5.41	CNN
	10^{-2}	11,737	0.939	7.93	MLP
ASCAD d0	0	63	0.006	1.71	CNN
	10^{-4}	264	0.026	3.28	MLP
	$3 \cdot 10^{-4}$	1,694	0.169	5.64	CNN
	10^{-3}	10,000	1.000	*8.00*	MLP
	10^{-2}	10,000	1.000	*8.00*	MLP
ASCAD d50	0	97	0.010	2.15	CNN
	10^{-4}	112	0.011	2.30	CNN
	$3 \cdot 10^{-4}$	141	0.014	2.56	CNN
	10^{-3}	815	0.082	4.68	CNN
	10^{-2}	10,000	1.000	*8.00*	CNN
ASCAD d100	0	98	0.010	2.16	CNN
	10^{-4}	109	0.011	2.28	CNN
	$3 \cdot 10^{-4}$	233	0.023	3.13	MLP
	10^{-3}	4,936	0.494	7.07	MLP
	10^{-2}	9,934	0.993	8.00	MLP

eight bits. The upper bounds on the three ASCAD variants range approximately from 1.5 to 2.3 bits, while those on AES-RD are 2.85, 5.34, and 6.80 bits for the MLP, CNN, and Transformer neighborhoods, respectively. The Transformer neighborhood has the largest upper bound among the three selected neighborhoods on AES-RD and ASCAD random key at this radius. These values describe the families covered at a smaller score radius, while $\varepsilon_{\mathrm{cal}} = 3 \cdot 10^{-4}$ remains the radius selected by the two calibration criteria.

Extending the bound to the full trained suite. Table 9 extends the evaluation to the union of score neighborhoods around all 80 MLP and CNN models on AES-RD and ASCAD d0, d50, and d100. This experiment is separate from the twelve selected-model curves and uses 12,500 certification traces for AES-RD and 10,000 for each ASCAD benchmark. It allocates failure probability $\delta/(80|\mathcal{E}|)$ to each binomial upper bound. For each radius, the table identifies a reference model whose neighborhood has the largest upper bound. The displayed rate is the fraction of certification traces satisfying that model's relaxed-margin condition. This fraction can exceed the model's classification accuracy because it includes some incorrect predictions.

At radius zero, the reported relaxed-margin counts equal the ordinary success counts. For ASCAD d0, both tables use 10,000 attack traces and count 63 successes, but the BI upper bounds are 1.675 bits for the finite suite and 1.709 bits for the local family. This difference follows from the confidence allocation. The suite interval allocates failure probability across the lower and upper bounds for all models, while the local-family calculation allocates it across upper bounds for all models and nine radii. Equal success counts therefore need not produce equal upper bounds.

The full-suite results also show how the upper bound changes with the permitted score deviation. On AES-RD, the BI upper bound remains near 5.4 bits

through $\varepsilon = 10^{-3}$ and then increases to 7.93 bits at 10^{-2}. On ASCAD d0, it increases from 1.71 bits at radius zero to 5.64 bits at the calibrated radius and reaches the trivial upper bound at 10^{-3}. Thus, expanding the attacker family through score neighborhoods can produce a substantially larger upper bound than certifying the fixed trained suite.

The reference model whose neighborhood determines the maximum upper bound can change as the radius increases. On AES-RD, a CNN determines this bound through $\varepsilon = 10^{-3}$, while an MLP determines it at $\varepsilon = 10^{-2}$ because more traces satisfy its relaxed-margin condition. These counts bound potential recovery within each neighborhood and need not be attained by a particular model. Taking the largest upper bound at each radius therefore provides a recovery guarantee for the entire union of score neighborhoods.

6 Conclusion, Limitations, and Future Work

Bounded Information (BI) uses empirical recovery outcomes on independent attack traces to provide a finite-sample upper confidence bound on the population recovery probability of a specified attacker family. It expresses this bound in bits relative to random guessing without estimating a high-dimensional leakage distribution. The finite-suite certificate bounds the maximum recovery probability among fixed trained models. The local-family certificate extends this guarantee to models whose normalized class scores differ from those of a reference model by at most a specified radius on every admissible trace. Experiments on multivariate AES traces show how the resulting upper bounds depend on the attack-set size and the attacker family. Reporting the empirical recovery probability together with its lower and upper confidence bounds distinguishes observed performance from statistical uncertainty. A separate CHES-CTF-2025 sweep shows numerical stability as the trace dimension increases to 7,000.

Limitations. A BI certificate applies to the specified attacker family. A small upper bound for a trained suite can therefore coexist with greater recovery success by an attacker outside that suite. Extending the guarantee to unrestricted Bayes vulnerability or mutual information requires Bayes-completeness, as stated in Section 3.5. Statistical uncertainty and attacker-family coverage consequently require different responses. Additional independent attack traces can improve statistical precision for a fixed family, whereas adding stronger models broadens coverage and may increase the certified upper bound. Similarly, increasing the score radius expands the local family and can increase its upper bound. More attack traces can reduce statistical uncertainty, but they need not eliminate the conservatism introduced by the relaxed-margin condition.

The guarantee also depends on the evaluation protocol. The models, preprocessing, score normalization, and radius grid must be fixed independently of the attack traces used for certification. These traces must be independent and identically distributed under the distribution to which the guarantee applies. Transfer learning-based attacks such as TranSCA [40] adapt a source model using labeled traces from a profiling device similar to the target device, while

UTLA [38] adapts across devices and leakage modalities using unlabeled target traces. For either approach, the evaluator requires independent labeled traces from the target distribution that were not used during adaptation. Separate calibration data can guide radius selection, but covering an additional model through the local-family guarantee requires a uniform score-distance bound over the admissible trace domain. Although the binomial confidence calculation has no explicit dependence on trace dimension, training and score-distance analysis can depend on trace dimension, architecture, and preprocessing. Here, trace dimension denotes the number of values in each trace supplied to the model, while attack-set size denotes the number of traces used for certification.

The single-trace certificate concerns exact recovery of the specified secret variable. It does not by itself determine the number of traces needed for key recovery because an attack can combine probability scores across traces. Our key-recovery experiments describe empirical performance, while direct population certification requires independent attack campaigns under Proposition 3. Applying this formulation to fixed CPA, DPA, or SASCA procedures remains an experimental direction.

Future work. Extending the evaluation to larger secret variables requires attention to attacker design and statistical precision. Moving from one byte to two increases the secret alphabet size from 256 to 65,536. When recovery remains close to random guessing, establishing a lower confidence bound above that baseline can require substantially more independent attack traces. Further work includes automating the derivation of tighter uniform score-distance bounds, particularly for attention-based models, and extending certification to changes in the device or measurement distribution. Investigating sharper confidence bounds, including PAC-Bayesian approaches [1], may also help reduce the statistical cost of covering larger attacker families.

Acknowledgment

The authors would like to thank the anonymous reviewers of IACR Asiacrypt 2026 for their constructive feedback, which greatly helped us in improving the final version of the manuscript. The authors would like to additionally acknowledge the Centre on Hardware-Security Entrepreneurship Research and Development (C-HERD), MeitY, Govt. of India; and Information Security Education and Awareness (ISEA) initiative, MeitY, Govt. of India; and Longview Philanthropy Inc. Ltd (UK) for partially funding this research.

References

1. Alquier, P.: User-friendly introduction to pac-bayes bounds. arXiv preprint arXiv:2110.11216 (2021), https://arxiv.org/abs/2110.11216
2. Batina, L., Bhasin, S., Jap, D., Picek, S.: CSI-NN: Reverse engineering of neural network architectures through electromagnetic side channel. In: 28th USENIX Security Symposium (USENIX Security 19). pp. 515–532 (2019), https://www.usenix.org/conference/usenixsecurity19/presentation/batina

3. Batina, L., Gierlichs, B., Prouff, E., Rivain, M., Standaert, F.X., Veyrat-Charvillon, N.: Mutual Information Analysis: A Comprehensive Study. Journal of Cryptology **24**, 269–291 (2011). https://doi.org/10.1007/s00145-010-9084-8
4. Battistello, A., Coron, J.S., Prouff, E., Zeitoun, R.: Horizontal Side-Channel Attacks and Countermeasures on the ISW Masking Scheme. In: Cryptographic Hardware and Embedded Systems (CHES) (2016). https://doi.org/10.1007/978-3-662-53140-2_2
5. Benadjila, R., Prouff, E., Strullu, R., Cagli, E., Dumas, C.: Deep learning for side-channel analysis and introduction to ascad database. Journal of Cryptographic Engineering **10**(2), 163–188 (2020). https://doi.org/10.1007/s13389-019-00220-8
6. Bhasin, S., Boyapally, H., Jap, D., Yap, T., Wu., Q.: GE Wars: The Deep Learning SCA battle. CHES Challenge 2025 (2025), https://pace-tl.gitbook.io/ches-challenge-2025
7. Bhasin, S., Bruneau, N., Danger, J.L., Guilley, S., Najm, Z.: Analysis and improvements of the dpa contest v4 implementation. In: International Conference on Security, Privacy, and Applied Cryptography Engineering. pp. 201–218. Springer (2014). https://doi.org/10.1007/978-3-319-12060-7_14
8. Bronchain, O., Hendrickx, J.M., Massart, C., Olshevsky, A., Standaert, F.X.: Leakage Certification Revisited: Bounding Model Errors in Side-Channel Security Evaluations. In: Advances in Cryptology – CRYPTO (2019). https://doi.org/10.1007/978-3-030-26948-7_25
9. Chari, S., Rao, J.R., Rohatgi, P.: Template attacks. In: International workshop on cryptographic hardware and embedded systems. pp. 13–28. Springer (2002). https://doi.org/10.1007/3-540-36400-5_3
10. Chérisey, C., Guilley, S., Rioul, O., Music, P.: Best Information is Most Successful: Mutual Information and Success Rate in Side-Channel Analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 49–79 (2019). https://doi.org/10.13154/tches.v2019.i2.49-79
11. Chowdhury, A., Roy, A., Brunetta, C., Oswald, E.: Leakage certification made simple. In: Annual International Cryptology Conference. pp. 427–460. Springer (2024). https://doi.org/10.1007/978-3-031-68391-6_13
12. Clopper, C.J., Pearson, E.S.: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26**(4), 404–413 (1934). https://doi.org/10.1093/biomet/26.4.404
13. Coron, J.S., Kizhvatov, I.: Analysis and improvement of the random delay countermeasure of ches 2009. In: International workshop on cryptographic hardware and embedded systems. pp. 95–109. Springer (2010). https://doi.org/10.1007/978-3-642-15031-9_7
14. Cover, T.M., Thomas, J.A.: Data compression. Elements of Information Theory pp. 103–158 (2005). https://doi.org/10.1002/047174882X.ch5
15. Duc, A., Dziembowski, S., Faust, S.: Unifying Leakage Models: From Probing Attacks to Noisy Leakage. In: Advances in Cryptology – EUROCRYPT (2014). https://doi.org/10.1007/978-3-642-55220-5_24
16. Duc, A., Faust, S., Standaert, F.X.: Making Masking Security Proofs Concrete: Or How to Evaluate the Security of Any Leaking Device. Journal of Cryptology **32**, 1263–1297 (2019). https://doi.org/10.1007/s00145-018-9277-0
17. Gao, W., Kannan, S., Oh, S., Viswanath, P.: Estimating mutual information for discrete-continuous mixtures. In: Advances in Neural Information Processing Systems. vol. 30 (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/ef72d53990bc4805684c9b61fa64a102-Paper.pdf

18. Hajra, S., Chowdhury, S., Mukhopadhyay, D.: Estranet: An efficient shift-invariant transformer network for side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems **2024**(1), 336–374 (2024). https://doi.org/10.46586/tches.v2024.i1.336-374
19. Hajra, S., Mukhopadhyay, D.: Multivariate Leakage Model for Improving Nonprofiling DPA on Noisy Power Traces. In: Information Security and Cryptology. pp. 325–342. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-12087-4_21
20. Heuser, A., Rioul, O., Guilley, S.: Good Is Not Good Enough: Deriving Optimal Distinguishers from Communication Theory. In: Cryptographic Hardware and Embedded Systems (CHES) (2014). https://doi.org/10.1007/978-3-662-44709-3_4
21. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American statistical association **58**(301), 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830
22. Ishai, Y., Sahai, A., Wagner, D.: Private Circuits: Securing Hardware against Probing Attacks. In: Advances in Cryptology – CRYPTO (2003). https://doi.org/10.1007/978-3-540-45146-4_27
23. Issa, I., Kamath, S., Wagner, A.B.: An operational measure of information leakage. In: 2016 Annual Conference on Information Science and Systems (CISS). pp. 234–239. IEEE (2016). https://doi.org/10.1109/CISS.2016.7460507
24. Ito, A., Ueno, R., Homma, N.: Perceived information revisited new metrics to evaluate success rate of side-channel attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems **2022**(4), 228–254 (2022). https://doi.org/10.46586/tches.v2022.i4.228-254
25. Ito, A., Ueno, R., Homma, N.: Perceived Information Revisited II: Theoretical Analysis of Deep-Learning Based Side-Channel Attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems (2025). https://doi.org/10.46586/tches.v2025.i1.450-474
26. Kim, J., Picek, S., Heuser, A., Bhasin, S., Hanjalic, A.: Make some noise. unleashing the power of convolutional neural networks for profiled side-channel analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 148–179 (2019). https://doi.org/10.13154/tches.v2019.i3.148-179
27. Mangard, S., Popp, T., Gammel, B.M.: Side-Channel Leakage of Masked CMOS Gates. In: CT-RSA (2005). https://doi.org/10.1007/978-3-540-30574-3_24
28. Massey, J.L.: Guessing and entropy. In: Proceedings of 1994 IEEE International Symposium on Information Theory. p. 204. IEEE (1994). https://doi.org/10.1109/ISIT.1994.394764
29. Masure, L., Cassiers, G., Hendrickx, J.M., Standaert, F.X.: Information Bounds and Convergence Rates for Side-Channel Security Evaluators. IACR Transactions on Cryptographic Hardware and Embedded Systems (3), 522–569 (2023). https://doi.org/10.46586/tches.v2023.i3.522-569
30. Masure, L., Strullu, R.: Side-channel analysis against anssi's protected aes implementation on arm: end-to-end attacks with multi-task learning. Journal of Cryptographic Engineering pp. 129–147 (2023). https://doi.org/10.1007/s13389-023-00311-7
31. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2018), https://mitpress.mit.edu/9780262039406/foundations-of-machine-learning/
32. Naghibijouybari, H., Neupane, A., Qian, Z., Abu-Ghazaleh, N.: Rendered insecure: GPU side channel attacks are practical. In: Proceedings of the 2018 ACM SIGSAC

Conference on Computer and Communications Security. pp. 2139–2153 (2018). https://doi.org/10.1145/3243734.3243831
33. Picek, S., Heuser, A., Jovic, A., Bhasin, S., Regazzoni, F.: The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 209–237 (2019). https://doi.org/10.13154/tches.v2019.i1.209-237
34. Picek, S., Perin, G., Mariot, L., Wu, L., Batina, L.: Sok: Deep learning-based physical side-channel analysis. ACM Computing Surveys **55**(11), 1–35 (2023). ht tps://doi.org/10.1145/3569577
35. Prouff, E., Rivain, M.: Masking against Side-Channel Attacks: A Formal Security Proof. In: Advances in Cryptology – EUROCRYPT (2013). https://doi.org/10.1007/978-3-642-38348-9_9
36. Prouff, E., Strullu, R., Benadjila, R., Cagli, E., Dumas, C.: Study of deep learning techniques for side-channel analysis and introduction to ASCAD database. IACR Cryptology ePrint Archive **2018**, 53 (2018), http://eprint.iacr.org/2018/053
37. Renauld, M., Standaert, F.X., Veyrat-Charvillon, N., Kamel, D., Flandre, D.: A Formal Study of Power Variability Issues and Side-Channel Attacks for Nanoscale Devices. In: Advances in Cryptology – EUROCRYPT (2011). https://doi.org/10.1007/978-3-642-20465-4_8
38. Roy, S.K., Zhang, Z., Ding, A.A., Fei, Y.: Attack from shadows: Unsupervised side-channel transfer learning across devices and modalities. In: Proceedings of the ACM Asia Conference on Computer and Communications Security (2026). https://doi.org/10.1145/3779208.3805969
39. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge university press (2014). https://doi.org/10.1017/CBO9781107298019
40. Thapar, D., Alam, M., Mukhopadhyay, D.: TranSCA: Cross-family profiled side-channel attacks using transfer learning on deep neural networks. Cryptology ePrint Archive, Paper 2020/1258 (2020), https://eprint.iacr.org/2020/1258
41. Timon, B.: Non-profiled deep learning-based side-channel attacks with sensitivity analysis. IACR Transactions on Cryptographic Hardware and Embedded Systems pp. 107–131 (2019). https://doi.org/10.13154/tches.v2019.i2.107-131
42. Xu, T., Roy, S.K., Ding, R., Ding, A.A., Fei, Y.: SLAC: Access-driven CPU-to-GPU side-channel attacks via system-level cache on Apple Silicon. arXiv preprint arXiv:2608.09075 (2026). https://doi.org/10.48550/arXiv.2608.09075
43. Zaid, G., Bossuet, L., Habrard, A., Venelli, A.: Methodology for efficient cnn architectures in profiling attacks. IACR Transactions on Cryptographic Hardware and Embedded Systems (1), 1–36 (Nov 2019). https://doi.org/10.13154/tches.v2020.i1.1-36

A Generative AI Disclosure

The paper was written directly by the authors. The first author used ChatGPT for basic editing, including feedback on technical writing, grammar, and spelling. Claude (Anthropic) was used to improve the aesthetics of the Matplotlib plots in the LaTeX manuscript.

B Proofs and Practical Material for the Bounded Information Certificate

This appendix describes how to compute the finite-suite and local-family BI certificates introduced in Sections 3 and 4. It then explains the role of Bayes-completeness and derives the singleton and finite-suite confidence bounds. Appendix C.2 gives the proof for the local family. We use the notation in Table 2, where $n_{\mathrm{te}} \geq 1$ is the number of independent attack traces used for certification, $M \geq 1$ is the number of fixed reference models, and $\delta \in (0, 1)$ is the failure probability. The trace dimension d denotes the number of values in each trace supplied to a model and is distinct from the attack-set size n_{te}.

B.1 Practical Procedure for Suite and Local-Family Certificates

Both certificates use a fixed suite of trained models and independent attack traces. The finite-suite certificate counts each model's correct predictions. The local-family certificate instead counts traces that satisfy a relaxed-margin condition for each reference model. This condition is necessary for a model within the specified score neighborhood to predict correctly, but it can also include traces that no model in the neighborhood recovers. A binomial upper confidence bound on the probability of this event therefore bounds recovery throughout the neighborhood. The following five steps describe the procedure, and Figure 3 illustrates the finite-suite calculation.

(1) Separate the data. Partition the data into profiling, validation, and attack sets. The attack trace-label pairs must be independent and identically distributed under the evaluation distribution. They must also be independent of all data used to develop the evaluated models and preprocessing. If calibration is used to select a score radius, reserve a separate calibration partition so that the traces used for certification remain independent of this selection.

(2) Train and select the models. Train candidate models on profiling traces and use validation traces to choose hyperparameters and select the final suite. Since these choices are independent of the certification traces, the finite-suite confidence calculation accounts for the M selected models. It requires no additional adjustment for candidates considered only during model development.

(3) Fix the evaluation. Specify the target variable, suite $\mathcal{H}_{\mathrm{tr}} = \{h_1, \ldots, h_M\}$, preprocessing, deterministic tie-breaking rule, and failure probability. For the local-family certificate, also fix the admissible trace domain, score representations, positive normalization constants A_j, and a finite nonempty grid $\mathcal{E}$ of radius vectors. Each vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_M)$ specifies a score radius for every reference model. All these choices must be independent of the certification traces. The score radii define the family being certified. To establish that a specified parameter perturbation produces a model in

this family, an architecture-specific analysis can bound its effect on scores uniformly over the admissible trace domain, as described in Section 4.

(4) Count successes and compute the bounds. Evaluate every suite member on the attack set. For the two-sided suite interval, compute a binomial lower confidence bound and an upper confidence bound from each success count S_j. Assign failure probability $\delta/(2M)$ to each one-sided bound. For the local family, use the relaxed-margin count $S_j^{\mathrm{loc}}(\varepsilon_j)$ in Equation (32) and assign failure probability $\delta/(M|\mathcal{E}|)$ to each upper confidence bound. Taking the maximum across reference models gives the following upper bounds on recovery probability.

$$\begin{aligned} p^{+}_{\mathrm{suite}} &= \max_{j \le M} U_{\mathrm{binom}}\left(S_j, n_{\mathrm{te}}, \frac{\delta}{2M}\right), \\ p^{+}_{\mathrm{loc}}(\varepsilon) &= \max_{j \le M} U_{\mathrm{binom}}\left(S_j^{\mathrm{loc}}(\varepsilon_j), n_{\mathrm{te}}, \frac{\delta}{M|\mathcal{E}|}\right). \end{aligned} \tag{46}$$

Each allocation provides confidence at least $1-\delta$ for its respective procedure. To obtain simultaneous coverage for both procedures at this confidence level, divide a total failure probability of δ between them and use the assigned value in each calculation. If the suite certificate requires only an upper bound, it can use δ/M for each model instead of the allocation for a two-sided interval.

(5) Report the result. For an upper bound $p^{+} \in [0,1]$ on recovery probability, report $\mathrm{BI}^{+} = \log_2(\max\{1, Kp^{+}\})$, following Definition 2. Include the probability bound, trace dimension, attack-set size, confidence allocation, and attacker family. For the finite suite, also report the maximum empirical recovery probability and the lower confidence bound on the maximum population recovery probability. For the local family, report the score radii, normalization constants, and relaxed-margin counts. Reporting these quantities together distinguishes empirical recovery performance from statistical uncertainty and the conservatism introduced by the relaxed-margin.

Remark 4 (Vacuous bounds, statistical margins, and decisions). An upper bound on recovery probability is vacuous when it equals one because it imposes no restriction beyond the range of a probability. The corresponding BI upper bound is $\mathrm{BI}^{+} = \log_2 K$. For a suite whose empirical recovery probability exceeds random guessing, the ratio R_{margin} exceeds one when the statistical margin is larger than the empirical advantage over $1/K$. This ratio describes statistical uncertainty, whereas a security decision depends on the confidence interval relative to a chosen recovery threshold. For example, a lower confidence bound above the threshold establishes vulnerability even when the upper bound is one, as explained in Section 3.4.

Additional independent attack traces can reduce statistical uncertainty for a fixed family. The local-family bound also includes the conservatism introduced by the relaxed-margin condition, which additional traces need not eliminate. In

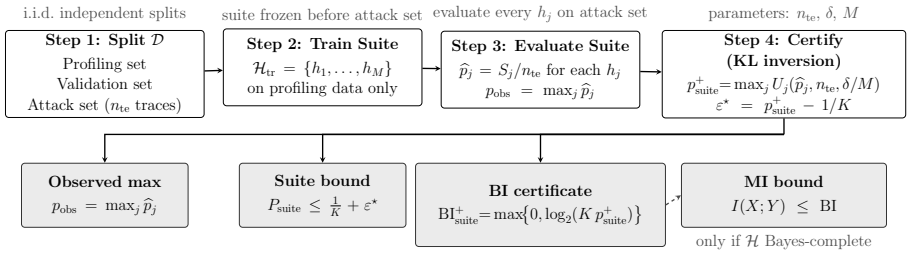


Fig. 3: The diagram illustrates finite-suite certification with failure probability δ/M assigned to each model's upper confidence bound. The procedure above also gives a two-sided suite interval using $\delta/(2M)$ for each one-sided bound and a separate local-family certificate using relaxed-margin counts. Interpreting BI as an upper bound on mutual information additionally requires Bayes-completeness.

particular, a relaxed-margin count equal to the attack-set size gives an upper bound of one. Recomputing the bound from the same outcomes cannot reduce this value. Additional traces may change the empirical fraction satisfying the condition, but they do not guarantee a useful upper bound for a broad score neighborhood.

Conditions for a valid evaluation. The simultaneous bound covers selection of the largest empirical recovery probability within a suite fixed independently of the certification traces. If those traces are instead used to construct or tune new models, the required independence no longer follows from the evaluation protocol. Increasing M after such adaptation does not generally restore the guarantee. The evaluator must then use new independent certification traces or a confidence bound that accounts for the adaptive procedure.

A proof of uniform score proximity also requires valid upper bounds on the norms used in the perturbation analysis. Power iteration can estimate a layer's spectral norm, but an estimate obtained after a finite number of iterations is not automatically an upper bound. The analysis must therefore use a justified upper bound, such as a matrix-norm bound known to dominate the spectral norm. Empirical score comparisons can guide radius calibration, but they do not establish score proximity over the full admissible trace domain.

B.2 Bayes Adversary and Bayes-Completeness

The Bayes attacker uses the true conditional distribution to predict a secret value with the highest posterior probability. With a fixed rule for resolving ties,

its prediction rule and recovery probability satisfy

$$\begin{aligned} h^\star(y) &\in \arg\max_{x\in\mathcal{X}} \Pr[X = x \mid Y = y], \\ P_{\text{guess}}(X \mid Y) &= \sup_h \Pr[h(Y) = X] \\ &= \mathbb{E}_Y\left[\max_{x\in\mathcal{X}} \Pr[X = x \mid Y]\right]. \end{aligned} \tag{47}$$

Here the supremum is taken over all measurable prediction rules. An attacker family $\mathcal{H}$ is Bayes-complete when its supremum recovery probability equals this unrestricted optimum. A family containing a Bayes rule satisfies this condition. However, equality of the suprema does not require a particular model in the family to attain the optimum.

Bayes-completeness allows a recovery certificate for the specified family to bound unrestricted guessing success. For uniform X, this also gives the mutual information bound in Section 3.5. Establishing this property requires a statement about the true distribution. For example, a template model implements a Bayes rule when its assumed distribution and parameters match the device distribution. Fitting a parametric model alone does not establish that its predictions are Bayes-optimal. When Bayes-completeness has not been established, BI certifies recovery within the specified family without automatically bounding all possible attackers or mutual information.

Remark 5 (Scope boundary). The single-trace experiments use models trained on labeled profiling data and evaluated on independent attack traces. The same Bernoulli argument applies to any prediction rule fixed independently of the certification traces. This can include a learning-based non-profiled procedure such as DDLA [41] when its use of data satisfies the same separation. For an attack that combines several traces, such as CPA, DPA, or SASCA, the recovery event must instead describe the outcome of the complete fixed procedure. Proposition 3 certifies this outcome using independent attack campaigns as statistical trials. Single-trace classification results cannot be substituted for these campaign outcomes because the two evaluations measure different recovery events.

B.3 Derivation of the Singleton Certificate

We condition on all profiling, validation, and other development data used to choose the attacker $\widehat{h}_{\text{sel}}$, together with any randomness used in training and selection. The selected prediction rule is then fixed, while the attack trace-label pairs $S_{\text{te}} = \{(X_i, Y_i)\}_{i=1}^{n_{\text{te}}}$ remain independent and identically distributed. The correctness indicators

$$Z_i := \mathbf{1}\{\widehat{h}_{\text{sel}}(Y_i) = X_i\} \tag{48}$$

are therefore independent Bernoulli variables with the same mean $p_{\text{succ}} := \Pr[\widehat{h}_{\text{sel}}(Y) = X]$. Their empirical mean is $\widehat{p}_{\text{succ}} = n_{\text{te}}^{-1} \sum_i Z_i$. Applying Ho-

effding's inequality [21] to this mean gives

$$\Pr\left[p_{\mathrm{succ}} \le \widehat{p}_{\mathrm{succ}} + \sqrt{\frac{\ln(1/\delta)}{2n_{\mathrm{te}}}}\right] \ge 1-\delta. \tag{49}$$

The upper bound may be clipped at one because a recovery probability cannot exceed one. This probability statement holds conditionally on the development data and model-selection randomness. Averaging over them therefore gives the same unconditional confidence level. Consequently, selecting one model using independent validation data requires no additional adjustment for the number of candidates considered during development.

The same reasoning applies to each suite member fixed independently of the attack traces. Appendix B.4 combines the individual confidence bounds to cover the entire suite. Theorem 1 uses the Bernoulli argument for relaxed-margin events to extend coverage to a local score family.

Binomial Confidence Bounds Hoeffding's inequality gives a statistical margin that depends only on the attack-set size and confidence level. Binomial confidence bounds also use the observed success count, allowing the statistical margin to depend on the empirical recovery probability. Let p denote the fixed attacker's population recovery probability and let $S := \sum_i Z_i$ be its success count. Then $S \sim \mathrm{Bin}(n_{\mathrm{te}}, p)$ and $\widehat{p} = S/n_{\mathrm{te}}$ is the empirical recovery probability. We describe two upper confidence bounds that are valid for finite attack sets. The main suite calculation uses the KL-binomial construction.

Clopper–Pearson bound. Inverting the binomial test gives the one-sided Clopper–Pearson upper confidence bound [12], defined as

$$p^{\mathrm{CP}}_{\mathrm{cert}} := \begin{cases} 1, & S = n_{\mathrm{te}}, \\ \mathrm{BetaInv}(1-\delta, S+1, n_{\mathrm{te}}-S), & S < n_{\mathrm{te}}, \end{cases} \tag{50}$$

where $\mathrm{BetaInv}(q, a, b)$ is the q-quantile of a $\mathrm{Beta}(a, b)$ distribution. This bound satisfies $\Pr[p \le p^{\mathrm{CP}}_{\mathrm{cert}}] \ge 1-\delta$ for every $p \in [0,1]$. The method is called exact because its validity does not rely on an asymptotic approximation. Since the binomial count is discrete, the actual coverage probability can exceed the specified confidence level.

KL-binomial bound. For $u, v \in [0,1]$, the binary Kullback–Leibler (KL) divergence is defined using natural logarithms as

$$D_{\mathrm{Ber}}(u\|v) := u \ln \frac{u}{v} + (1-u) \ln \frac{1-u}{1-v}, \tag{51}$$

with the usual limiting conventions at zero and one. Inverting the Bernoulli Chernoff bound gives the upper confidence bound

$$p^{\mathrm{KL}}_{\mathrm{cert}} := \sup\left\{q \in [\widehat{p}, 1] : n_{\mathrm{te}} D_{\mathrm{Ber}}(\widehat{p}\|q) \le \ln(1/\delta)\right\}. \tag{52}$$

To establish coverage, fix the population recovery probability p. The bound fails only if $p > p^{\mathrm{KL}}_{\mathrm{cert}}$, which requires $\widehat{p} < p$ and $n_{\mathrm{te}} D_{\mathrm{Ber}}(\widehat{p}\|p) > \ln(1/\delta)$. These inequalities describe success counts sufficiently far below their population mean. The Chernoff bound assigns probability at most δ to this lower-tail event. Therefore, $\Pr[p \le p^{\mathrm{KL}}_{\mathrm{cert}}] \ge 1 - \delta$.

For $0 < \widehat{p} < 1$, the divergence increases with q on $(\widehat{p}, 1)$, so the upper bound can be computed by one-dimensional bisection. When $S = 0$, it is $p^{\mathrm{KL}}_{\mathrm{cert}} = 1 - \delta^{1/n_{\mathrm{te}}}$. When $S = n_{\mathrm{te}}$, the upper bound is one. The Clopper–Pearson method gives the same upper bounds for these two counts.

Either method also gives a lower confidence bound on recovery probability by applying its upper-bound construction to the number of incorrect predictions. Let $U_{\mathrm{binom}}(S, n, \alpha)$ denote the chosen upper confidence bound for S successes among n trials, with failure probability α. The corresponding lower confidence bound is

$$L_{\mathrm{binom}}(S, n, \alpha) := 1 - U_{\mathrm{binom}}(n - S, n, \alpha). \tag{53}$$

For the KL method, this lower bound equals the infimum over $q \in [0, \widehat{p}]$ used in Equation (15). Each one-sided bound fails with probability at most α.

Converting the upper bound to BI. Either method provides an upper confidence bound $p_{\mathrm{cert}} \in [0, 1]$ on population recovery probability. Following the nonnegative reporting convention in Definition 2, we set $\overline{\varepsilon}^{\star} = \max\{0, p_{\mathrm{cert}} - 1/K\}$ and report

$$\mathrm{BI}^{+} = \log_2(1 + K\overline{\varepsilon}^{\star}) = \log_2(\max\{1, K p_{\mathrm{cert}}\})\,. \tag{54}$$

The conversion is nondecreasing in the probability bound and therefore preserves the confidence guarantee. Binomial bounds can reduce the statistical margin relative to Hoeffding's inequality when empirical recovery probability is low. Their validity does not require a strictly smaller margin for every possible success count. In particular, when every attack trace is classified correctly, both binomial methods and the clipped Hoeffding bound give an upper bound of one.

B.4 Proof of Proposition 2: Finite-Suite Certificate

We condition on the suite chosen independently of the attack traces, together with its development data and training randomness. For each fixed model h_j, the success count S_j is binomial with parameters n_{te} and $p_j := \Pr[h_j(Y) = X]$. Each one-sided confidence bound uses failure probability $\alpha = \delta/(2M)$ and fails with probability at most α. A union bound over the lower and upper bounds for all M models therefore gives the following inequalities simultaneously for every j with probability at least $1 - \delta$.

$$L_{\mathrm{binom}}(S_j, n_{\mathrm{te}}, \delta/(2M)) \le p_j \le U_{\mathrm{binom}}(S_j, n_{\mathrm{te}}, \delta/(2M)). \tag{55}$$

The union bound remains valid when different models make dependent predictions on the same traces. Averaging over the development data and training randomness removes the conditioning without changing the confidence level.

On this simultaneous confidence event, taking the maximum of the lower bounds and the maximum of the upper bounds yields

$$\max_j L_{\mathrm{binom}}(S_j, n_{\mathrm{te}}, \delta/(2M)) \le \max_j p_j \le \max_j U_{\mathrm{binom}}(S_j, n_{\mathrm{te}}, \delta/(2M)). \quad (56)$$

The middle quantity is the maximum population recovery probability within the suite. The inequalities therefore prove the confidence interval in Proposition 2. Using the KL-binomial lower and upper bounds gives the construction in Equation (14)-(15). The model with the highest empirical recovery probability need not have the highest population recovery probability. Simultaneous coverage of all suite members accounts for this uncertainty before their bounds are combined.

For comparison, applying Hoeffding's inequality separately to each tail with the same failure probability gives

$$p_j \in \left[\widehat{p}_j - \sqrt{\frac{\ln(2M/\delta)}{2n_{\mathrm{te}}}},\ \widehat{p}_j + \sqrt{\frac{\ln(2M/\delta)}{2n_{\mathrm{te}}}}\right] \cap [0,1] \qquad \text{for every } j \quad (57)$$

on an event of probability at least $1-\delta$. The suite size enters through $\ln(2M/\delta)$, which appears inside the square root in the Hoeffding bound. In the KL-binomial construction, the same term sets the divergence threshold. Neither confidence calculation has an explicit dependence on trace dimension, although trace dimension can affect the models' success counts.

The local-family certificate uses the same probability argument after replacing correctness indicators with relaxed-margin indicators and allocating failure probability across the radius grid. At radius zero, each model in a reference neighborhood has the same normalized score function as its reference model and therefore makes the same predictions under the common deterministic tie-breaking rule. However, the relaxed-margin event $\rho_j \ge 0$ includes traces on which the true class shares the maximum score even if the tie-breaking rule selects another class. The relaxed-margin count can therefore exceed the ordinary success count even at radius zero. Absence of such ties on the evaluated traces is sufficient for the counts to agree. When the counts agree, using the same attack traces, binomial method, and failure probability for each upper bound gives equal suite and zero-radius upper bounds, as discussed in Remark 3.

C Local-Family Proofs, Bayes Interpretation, and Composition

The local-family certificate uses a deterministic score-distance condition to identify which traces a model in a reference neighborhood can classify correctly. A binomial upper bound then bounds the probability of this event using independent attack traces. We first prove this certificate and derive sufficient score-distance bounds for MLPs, CNNs, and Transformers. We then state the additional assumptions needed to bound Bayes-optimal recovery and mutual information,

before considering attacks that combine evidence across targets and traces. The local-family recovery certificate applies without these additional assumptions.

C.1 Definitions and Notation

We represent a trace as $Y \in \mathbb{R}^{T \times C_0}$, where T is the number of temporal sampling positions and C_0 is the number of measurement channels. The trace dimension is $d = TC_0$, which counts all values in the trace supplied to the model. This differs from the attack-set size n_{te}, which counts the independent traces used for certification. The deterministic score-distance bounds apply on an admissible domain $\mathcal{Y}_B$ satisfying $\|\mathrm{vec}(Y)\|_2 \le B_Y$. This domain must contain the traces supported by the evaluation distribution. A maximum norm observed on a finite dataset alone does not establish the required uniform norm bound.

For a temporal window length $1 \le \tau \le T$, we define $P_t(Y) = Y_{t:t+\tau-1,:}$ for $1 \le t \le N_\tau$, where $N_\tau = T - \tau + 1$. Each patch contains τ consecutive sampling positions across all channels and therefore has dimension τC_0. For a stack of L stride-one valid convolutions without dilation, the receptive field spans $R_L = 1 + \sum_{\ell=1}^{L}(k_\ell - 1)$ sampling positions, where k_ℓ is the kernel width of layer ℓ. An output unit can directly combine values separated by $\tau - 1$ temporal indices only if its receptive field includes both positions. A subsequent aggregation stage may combine information from several units, so the complete network requires a separate analysis of that stage.

All vector norms below are Euclidean unless an infinity norm is specified. Norms of traces and patches refer to their vectorized representations. The operator norm of a weight matrix bounds the amplification caused by its linear transformation and satisfies $\|W\|_{\mathrm{op}} \le \|W\|_F$, where $\|W\|_F$ is the Frobenius norm. An affine layer followed by a one-Lipschitz activation has Lipschitz constant at most $\|W\|_{\mathrm{op}}$. Its output norm also depends on the bias and the activation's value at zero, so these terms must be included when bounding intermediate activations. Throughout the analysis, we use the score normalization $\bar{s}_h^{(j)} = s_h / A_j$ from Definition 4, with the same positive constant A_j for the candidate and reference models in each comparison.

C.2 Proof of the Local-Family Margin Certificate

A score radius limits how much each class score can change relative to the reference model. Consequently, a model in the neighborhood can correct a reference error only if the true-class score is sufficiently close to the highest competing score. The proof first expresses this restriction as a relaxed-margin event and then bounds its probability simultaneously over the reference models and radius grid.

Proof (Proof of Theorem 1). We first show that every correct prediction in a reference neighborhood satisfies the relaxed-margin condition. Fix a reference model h_j, a radius ε_j, and a model $h \in \mathcal{H}_j^{\mathrm{loc}}(\varepsilon_j)$. We suppress the superscript (j) on the normalized scores during this comparison. If $h(y) = x$, then $\bar{s}_{h,x}(y) \ge$

$\bar{s}_{h,a}(y)$ for every competing class a. The uniform score-distance bound therefore gives

$$\bar{s}_{h_j,x}(y) \geq \bar{s}_{h,x}(y) - \varepsilon_j \geq \bar{s}_{h,a}(y) - \varepsilon_j \geq \bar{s}_{h_j,a}(y) - 2\varepsilon_j. \tag{58}$$

Thus, the reference model's true-class margin satisfies $\rho_j(x, y) \geq -2\varepsilon_j$. The factor of two accounts for an increase of ε_j in the true-class score and a decrease of ε_j in a competing score. Every admissible trace-label pair therefore satisfies

$$\mathbf{1}\{h(y) = x\} \leq \mathbf{1}\{\rho_j(x, y) \geq -2\varepsilon_j\}. \tag{59}$$

Taking expectations gives the following upper bound on the recovery probability of every model in neighborhood j.

$$\Pr[h(Y) = X] \leq q_j(\varepsilon_j), \qquad q_j(\varepsilon_j) := \Pr[\rho_j(X, Y) \geq -2\varepsilon_j]. \tag{60}$$

We next bound the relaxed-margin probability using independent attack traces. We condition on the suite, preprocessing, score representations, normalization constants, and radius grid, which are all fixed independently of the certification data. The relaxed-margin count then follows a binomial distribution with n_{te} trials and success probability $q_j(\varepsilon_j)$. Appendix B.3 therefore gives

$$q_j(\varepsilon_j) \leq U_{\mathrm{binom}}\left(S_j^{\mathrm{loc}}(\varepsilon_j), n_{\mathrm{te}}, \frac{\delta}{M|\mathcal{E}|}\right) \tag{61}$$

except with probability at most $\delta/(M|\mathcal{E}|)$. A union bound over the M references and $|\mathcal{E}|$ radius vectors makes all these statements hold with probability at least $1 - \delta$. Independence between models or radii is not needed.

On this simultaneous confidence event, each model in $\mathcal{H}_{\mathrm{loc}}(\varepsilon)$ belongs to at least one reference neighborhood, so its recovery probability satisfies that neighborhood's upper bound. Taking the supremum over the family gives

$$\sup_{h \in \mathcal{H}_{\mathrm{loc}}(\varepsilon)} \Pr[h(Y) = X] \leq \max_{j \leq M} U_{\mathrm{binom}}\left(S_j^{\mathrm{loc}}(\varepsilon_j), n_{\mathrm{te}}, \frac{\delta}{M|\mathcal{E}|}\right). \tag{62}$$

The same event covers every radius vector in the fixed grid. Averaging over the independent development data removes the conditioning and proves the theorem.

The relaxed-margin count includes every trace that a model in the neighborhood could classify correctly, although a single model need not achieve all of these successes. At radius zero, the count also includes traces whose true class shares the largest score with another class. The fixed tie-breaking rule may classify some of these traces incorrectly, so the relaxed count equals the ordinary success count under an additional condition such as the absence of these ties. When the counts agree, equality with the finite-suite upper bound additionally requires the same attack traces and the same failure probability allocated to each upper bound, as explained in Remark 3.

C.3 Deterministic Score-Distance Bounds

The previous proof certifies a family defined by a uniform score-distance condition. To establish that specified parameter perturbations produce models in this family, we bound their effect on the output scores over the admissible trace domain. An upper bound $\sup_{Y \in \mathcal{Y}_B} \|s_h(Y) - s_{h_j}(Y)\|_\infty \le \Delta_j$ gives the normalized radius $\varepsilon_j = \Delta_j / A_j$. The following lemmas provide sufficient bounds for three architectures. Applying a lemma requires verifying its assumptions and constants for the evaluated model.

Lemma 2 (MLP score-distance bound). *Consider an L-layer MLP. The reference and candidate models use the same fixed one-Lipschitz activation σ_ℓ in each layer. The final activation may be the identity. Assume that the vectorized input satisfies $\|Y\|_2 \le B_Y$ and that the reference weights satisfy $\|W_j^{(\ell)}\|_{\mathrm{op}} \le B_{j,\ell}$. Suppose the weight and bias perturbations satisfy*

$$\|W^{(\ell)} - W_j^{(\ell)}\|_{\mathrm{op}} \le r_\ell, \qquad \|b^{(\ell)} - b_j^{(\ell)}\|_2 \le \eta_\ell. \tag{63}$$

Let $c_\ell := \|\sigma_\ell(0)\|_2$. We define an upper bound R_ℓ on the candidate activation norm after each layer by the recursion

$$R_0 = B_Y, \qquad R_\ell = (B_{j,\ell} + r_\ell) R_{\ell-1} + \|b_j^{(\ell)}\|_2 + \eta_\ell + c_\ell. \tag{64}$$

These activation bounds give the following uniform upper bound on the score difference.

$$\begin{aligned} \sup_{Y \in \mathcal{Y}_B} \|s_h(Y) - s_{h_j}(Y)\|_\infty &\le \Delta_j^{\mathrm{MLP}}(r, \eta), \\ \Delta_j^{\mathrm{MLP}}(r, \eta) &:= \sum_{\ell=1}^{L} (r_\ell R_{\ell-1} + \eta_\ell) \prod_{q>\ell} B_{j,q}. \end{aligned} \tag{65}$$

Consequently, $\Delta_j^{\mathrm{MLP}}(r, \eta)/A_j$ is a valid normalized radius.

Proof. One-Lipschitz continuity gives $\|\sigma_\ell(z)\|_2 \le \|z\|_2 + c_\ell$, which proves the activation recursion. Let D_ℓ bound the Euclidean difference between candidate and reference activations after layer ℓ, with $D_0 = 0$. Adding and subtracting the reference weight applied to the candidate activation gives

$$D_\ell \le B_{j,\ell} D_{\ell-1} + r_\ell R_{\ell-1} + \eta_\ell. \tag{66}$$

Because both networks use the same activation, its offset contributes to the activation norm bound but introduces no separate term in the difference recursion. Expanding this recursion from $D_0 = 0$ gives the sum in Equation (65). The score difference in infinity norm is at most its Euclidean norm, which proves the stated upper bound.

The offset term vanishes when $\sigma_\ell(0) = 0$. If both models also have zero biases, the recursion reduces to the product bound in Section 4. Omitting the offset therefore requires the activation to map zero to zero in addition to being one-Lipschitz.

Lemma 3 (CNN score-distance bound with normalized pooling). *Consider a CNN whose score vector is computed by a shared patch network, normalized average pooling, and an affine output layer as follows.*

$$s_W(Y) = W_{\text{out}} \left(\frac{1}{N_\tau} \sum_{t=1}^{N_\tau} \phi_W(P_t(Y)) \right) + b_{\text{out}}. \tag{67}$$

Assume that every admissible patch has vectorized norm at most B_{patch} and that the shared patch network satisfies the assumptions of Lemma 2. Its proof gives a uniform upper bound Δ_{patch} on the Euclidean difference between candidate and reference patch outputs and an upper bound R_{patch} on the candidate output norm. Let B_{out} bound the operator norm of the reference output weight matrix. Let r_{out} and η_{out} bound the output weight perturbation in operator norm and the output bias perturbation in Euclidean norm, respectively. The complete score difference then satisfies

$$\begin{aligned} \sup_{Y \in \mathcal{Y}_B} \|s_W(Y) - s_{W_j}(Y)\|_\infty &\le \Delta_j^{\text{CNN}}, \\ \Delta_j^{\text{CNN}} &:= B_{\text{out}} \Delta_{\text{patch}} + r_{\text{out}} R_{\text{patch}} + \eta_{\text{out}}. \end{aligned} \tag{68}$$

Dividing by A_j gives a valid normalized radius.

Proof. Normalized averaging preserves both uniform patch bounds. In particular,

$$\left\| \frac{1}{N_\tau} \sum_t \big(\phi_W(P_t(Y)) - \phi_{W_j}(P_t(Y)) \big) \right\|_2 \le \frac{1}{N_\tau} \sum_t \Delta_{\text{patch}} = \Delta_{\text{patch}}. \tag{69}$$

The averaged candidate output also has norm at most R_{patch}. Applying the affine-layer difference argument from the MLP proof gives three contributions. The reference output weights amplify the pooled feature difference by at most $B_{\text{out}} \Delta_{\text{patch}}$, the weight perturbation contributes at most $r_{\text{out}} R_{\text{patch}}$, and the bias perturbation contributes at most η_{out}. Their sum proves the stated upper bound.

Normalized averaging introduces no additional multiplicative factor for the number of patches. Dependence on trace dimension can still enter through the patch domain, parameter bounds, or output layer. Flattening, unnormalized sums, and dense output layers with separate weights for each position require aggregation bounds appropriate to those operations.

Lemma 4 (Fixed-length Transformer score-distance bound). *Fix the tokenization, sequence length, preprocessing, attention pattern, and network structure. Write a reference Transformer as $G_j \circ F_{j,B} \circ \cdots \circ F_{j,1}$ and a candidate as $G \circ F_B \circ \cdots \circ F_1$. The complete output head G_j includes any final pooling and score transformation.*

Assume the activation domains cover every candidate and reference state arising in the layerwise comparison. On these domains, reference block b is

$L_{j,b}$-*Lipschitz and* $\|F_b(z) - F_{j,b}(z)\| \le \Delta_{j,b}(r_b)$. *Assume also that* G_j *is* $L_{j,\mathrm{out}}$-*Lipschitz from the final activation norm to the score infinity norm, and that* $\|G(z) - G_j(z)\|_\infty \le \Delta_{j,\mathrm{head}}$ *on the domain of candidate final activations. These assumptions give the following uniform score-distance bound.*

$$\sup_{Y \in \mathcal{Y}_B} \|s_h(Y) - s_{h_j}(Y)\|_\infty \le \Delta_j^{\mathrm{Tr}}(r),$$
$$\Delta_j^{\mathrm{Tr}}(r) := \Delta_{j,\mathrm{head}} + L_{j,\mathrm{out}} \sum_{b=1}^{B} \Delta_{j,b}(r_b) \prod_{q>b} L_{j,q}. \tag{70}$$

The normalized radius is $\Delta_j^{\mathrm{Tr}}(r)/A_j$.

Proof. We first bound the activation differences at each block and then at the output head. Let D_b bound the difference between candidate and reference activations after block b. Identical input preprocessing gives $D_0 = 0$. Adding and subtracting the reference block evaluated on the candidate input gives

$$D_b \le \Delta_{j,b}(r_b) + L_{j,b} D_{b-1}. \tag{71}$$

Unrolling this inequality bounds the final activation difference by the sum in Equation (70). Adding and subtracting the reference head evaluated on the candidate final activation then bounds the score difference by $\Delta_{j,\mathrm{head}} + L_{j,\mathrm{out}} D_B$.

The block constants must account for attention, residual connections, normalization, and feed-forward transformations. The output head contributes both through its amplification of earlier activation differences and through changes to its own parameters. Here, sequence length denotes the number of tokens produced by the fixed tokenization. Fixing this number allows the bound to be evaluated for that architecture, but does not establish that the constants remain unchanged when the sequence length or trace dimension changes.

Remark 6 (Score-distance bounds and empirical diagnostics). A fixed radius grid defines score neighborhoods that can be certified by the margin theorem. A parameter perturbation bound may nevertheless be too conservative to establish that a particular model belongs to one of these neighborhoods. Comparing scores on finitely many traces can guide radius calibration, but the largest observed difference may be smaller than the supremum over $\mathcal{Y}_B$. Establishing membership therefore requires a uniform bound on that domain.

Remark 7 (Optional capacity diagnostics). For a nonzero matrix W, its stable rank is $\|W\|_F^2 / \|W\|_{\mathrm{op}}^2$, which is at most its rank. These quantities and products of layer norms describe properties of the model, while the binomial upper confidence bound is computed from the success count, attack-set size, and confidence allocation. For layer widths $700 \to 64 \to 32 \to 16 \to 256$, the upper bounds on the individual weight-matrix ranks sum to $64 + 32 + 16 + 16 = 128$. The stable rank of the full convolution operator can grow with trace dimension, so a bound for the shared patch map requires further analysis before it can be applied to that operator.

C.4 Bayes-Completeness and the Mutual-Information Bound

Bayes-completeness means that the supremum recovery probability over the attacker family equals the unrestricted Bayes-optimal recovery probability. This property allows a family recovery certificate to bound unrestricted recovery and, through a lower bound on conditional entropy, mutual information. We first derive this conversion and then state conditions under which a family with shared patch processing and additive pooling can approach Bayes-optimal recovery. Applying the conversion to a score neighborhood around trained models requires Bayes-completeness of that specific neighborhood.

Proof (Proof of Lemma 1). We derive an entropy lower bound from the guessing probability and then apply the family certificate on its confidence event. For a trace y, let $p_y(x) = \Pr[X = x \mid Y = y]$ denote the posterior probability of class x. The Shannon entropy of this posterior satisfies

$$H(p_y) \geq -\log_2 \max_x p_y(x). \tag{72}$$

Taking expectations and applying Jensen's inequality to the convex function $-\log_2$ gives

$$\begin{aligned} H(X \mid Y) &= \mathbb{E}_Y H(p_Y) \\ &\geq \mathbb{E}_Y[-\log_2 \max_x p_Y(x)] \\ &\geq -\log_2 \mathbb{E}_Y[\max_x p_Y(x)]. \end{aligned} \tag{73}$$

The final expectation is the unrestricted guessing probability. Since X is uniform on K classes, substituting $H(X) = \log_2 K$ into $I(X;Y) = H(X) - H(X \mid Y)$ gives

$$I(X;Y) \leq \log_2\big(K P_{\text{guess}}(X \mid Y)\big). \tag{74}$$

Bayes-completeness identifies this guessing probability with $P^{\mathcal{H}}_{\text{guess}}$. Because unrestricted guessing succeeds with probability at least $1/K$, the population BI definition in Section 3 equals $\log_2(K P^{\mathcal{H}}_{\text{guess}})$. On the certificate's confidence event, we apply the recovery upper bound and restrict it to at most one to obtain

$$I(X;Y) \leq \mathrm{BI}^{\mathcal{H}}_{\text{pop}}(X;Y) \leq \log_2(1 + K\overline{\varepsilon}^{\star}), \tag{75}$$

where $\overline{\varepsilon}^{\star}$ is the clipped advantage in Definition 2. All conversions are deterministic on that event, so the confidence remains $1 - \delta$.

Without Bayes-completeness, the certificate bounds recovery within the specified family. Extending it to the unrestricted guessing probability in Equation (74) requires control of the gap in recovery probability, which cannot be replaced directly by an approximation error measured in score units. For nonuniform X, the entropy argument gives $I(X;Y) \leq H(X) + \log_2 P_{\text{guess}}(X \mid Y)$. The expression using $\log_2 K$ follows from the uniform-target setting adopted in the main text.

Assumption 1 (Additive local Bayes scores) *Let the patch support* $\mathcal{P}_\tau = \{P_t(Y) : Y \in \mathcal{Y}, 1 \le t \le N_\tau\}$ *be compact. There is a continuous map* $\psi : \mathcal{P}_\tau \to \mathbb{R}^K$ *such that a Bayes prediction can be formed from*

$$s^\star(Y) = \frac{1}{N_\tau} \sum_{t=1}^{N_\tau} \psi(P_t(Y)), \qquad h^\star(Y) \in \arg\max_x s_x^\star(Y). \tag{76}$$

We resolve ties using a fixed class-priority rule and define the Bayes-score margin by

$$m^\star(Y) := s^\star_{h^\star(Y)}(Y) - \max_{a \neq h^\star(Y)} s_a^\star(Y). \tag{77}$$

Assume that one of two conditions holds. Under condition (a), there is a constant $\gamma_0 > 0$ *such that* $m^\star(Y) \ge \gamma_0$ *almost surely. Under condition (b), the margin satisfies* $\Pr[m^\star(Y) = 0] = 0$.

This assumption states that averaging one shared patch-score function across positions suffices to make a Bayes-optimal prediction. Establishing it requires a property of the leakage distribution and score representation beyond the presence of repeated operations. Condition (a) provides a positive lower bound on the margin, whereas condition (b) permits arbitrarily small margins as long as exact ties have probability zero.

Proposition 4 (Bayes-completeness of pooled local families). *Suppose Assumption 1 holds. Let* $\mathcal{H}_\tau$ *contain classifiers whose pooled scores have the form* $N_\tau^{-1} \sum_t \phi(P_t(Y))$. *Suppose the allowable patch maps* ϕ *approximate* ψ *arbitrarily closely in the uniform norm on* $\mathcal{P}_\tau$. *Any output transformation must preserve class ordering, and all models must use the fixed tie rule. The family is then Bayes-complete, so its supremum recovery probability satisfies*

$$P_{\text{guess}}^{\mathcal{H}_\tau}(X \mid Y) = P_{\text{guess}}(X \mid Y). \tag{78}$$

Under condition (a), some member agrees with the Bayes prediction almost surely. Under condition (b), the family can approach Bayes success arbitrarily closely, without necessarily attaining it.

Shorter temporal windows can be insufficient for the additive score form in Definition 3. For every window length $\tau' < \tau$, *there is a binary-target distribution satisfying Assumption 1 at window length* τ *for which the supremum recovery probability of the family with window length* τ' *is strictly below the Bayes-optimal recovery probability.*

The approximation hypothesis must be established for the specified architecture class. In particular, the allowable patch networks must approximate the shared map ψ to any desired uniform accuracy. Fixed widths or parameter constraints can prevent this approximation even when the architecture is a CNN, MLP, or Transformer.

Remark 8 (Representability and learning). The proposition concerns the supremum recovery probability over the entire family. A model obtained through finite training may achieve a smaller recovery probability. Experiments that vary network width or temporal window length show changes in empirical performance, but do not generally identify the Bayes optimum, the population optimum at each width, or the optimization gap. The decomposition in Appendix C.5 expresses these gaps in recovery probability and distinguishes them from certification uncertainty.

C.5 Proof of Proposition 4

Uniform approximation and prediction agreement. For any $a > 0$, choose a patch map ϕ_a with $\sup_{p \in \mathcal{P}_\tau} \|\phi_a(p) - \psi(p)\|_\infty \le a$. Normalized averaging gives the same bound on the pooled scores for every admissible trace. The resulting prediction h_a agrees with $h^\star$ whenever $m^\star(Y) > 2a$, because no competing score can close a larger gap.

Under condition (a), we choose $a < \gamma_0/2$. The predictions then agree almost surely, so the model attains the Bayes-optimal recovery probability. Under condition (b), the absence of ties implies that the disagreement probability satisfies

$$\Pr[h_a(Y) \neq h^\star(Y)] \le \Pr[m^\star(Y) \le 2a] \longrightarrow 0 \qquad \text{as } a \downarrow 0. \tag{79}$$

The difference between the two success probabilities is at most their disagreement probability. Hence $\Pr[h_a(Y) = X] \to P_{\mathrm{guess}}(X \mid Y)$, which proves equality of suprema. The argument does not assert exact agreement for any particular positive approximation error under condition (b).

A counterexample for shorter temporal windows. We construct a parity target whose exact recovery requires combining all τ trace coordinates under additive pooling. Let $T = \tau$, $C_0 = 1$, and Y be uniform on $\{-1, +1\}^\tau$, so the trace dimension is $d = \tau$. Define $\chi(Y) = \prod_{i=1}^{\tau} Y_i$ and $X = \mathbf{1}\{\chi(Y) = 1\}$. The complete trace determines X, so the Bayes-optimal recovery probability is one. A window of length τ contains the entire trace, and the scores $s_1^\star = \chi$ and $s_0^\star = -\chi$ satisfy the assumption with margin two. The patch support is finite and therefore compact, and the patch map is continuous on that support.

For a binary model with window length $\tau' < \tau$, let $f(Y)$ be its pooled class-1 score minus its pooled class-0 score before any common output transformation that preserves class ordering. Each additive term depends on at most τ' coordinates, so its multilinear expansion on the Boolean cube has degree at most τ'. Consequently, f is orthogonal to the parity function under the uniform distribution, which gives

$$\mathbb{E}[\chi(Y) f(Y)] = 0. \tag{80}$$

To relate this identity to classification, suppose the model recovered parity correctly at every point. Then $\chi(Y) f(Y) \ge 0$ everywhere, with strict positivity at each point whose correct class is not favored by the fixed tie rule. That class has positive probability, which would make the expectation strictly positive and

contradict the displayed identity. Every such classifier therefore makes at least one error among the 2^τ equally likely inputs, giving an upper bound of $1 - 2^{-\tau}$ on its recovery probability. This uniform upper bound also applies to the supremum over the family. The counterexample depends on additive pooling followed by an argmax and does not establish the same restriction for arbitrary nonlinear aggregation of patch outputs.

Separating approximation, estimation, and optimization. Let $p(h) = \Pr[h(Y) = X]$, $p^\star = P_{\mathrm{guess}}(X \mid Y)$, and $\mathcal{H}_{\tau,W}$ be a fixed width-restricted class containing the trained model $\widehat{h}$. Define its population approximation gap by

$$\varepsilon_{\mathrm{approx}}(W) := p^\star - \sup_{h \in \mathcal{H}_{\tau,W}} p(h). \tag{81}$$

This approximation gap measures the difference in recovery probability between the Bayes optimum and the best value available within the specified class. To describe learning from profiling traces, let $\widehat{p}_{\mathrm{pr}}(h)$ denote the empirical recovery probability on those traces. We define the estimation and optimization terms by

$$\begin{aligned} \varepsilon_{\mathrm{est}} &:= 2 \sup_{h \in \mathcal{H}_{\tau,W}} |p(h) - \widehat{p}_{\mathrm{pr}}(h)|, \\ \varepsilon_{\mathrm{opt}} &:= \sup_{h \in \mathcal{H}_{\tau,W}} \widehat{p}_{\mathrm{pr}}(h) - \widehat{p}_{\mathrm{pr}}(\widehat{h}). \end{aligned} \tag{82}$$

Bounding the population supremum by the empirical supremum plus its uniform deviation, and applying the same deviation bound to the trained model, yields

$$p^\star - p(\widehat{h}) \le \varepsilon_{\mathrm{approx}}(W) + \varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{opt}}. \tag{83}$$

This inequality holds deterministically for the defined quantities. A probabilistic upper bound on $\varepsilon_{\mathrm{est}}$ requires a uniform bound on the difference between empirical and population recovery probabilities over the learning class. The term $\varepsilon_{\mathrm{opt}}$ measures a gap in empirical recovery probability, which may differ from the training objective. Experiments that vary network width do not automatically measure either term. The binomial confidence interval for a fixed trained model instead bounds uncertainty about $p(\widehat{h})$ using independent certification traces. The decomposition therefore separates these sources of error without specifying a universal rate of convergence with network width.

C.6 Local-Family MI Bound and Interpretation

A broad architecture class can be Bayes-complete even when a score neighborhood around a trained suite has a smaller supremum recovery probability. The following corollary therefore requires Bayes-completeness of the specific local family being certified. If its approximation gap is nonzero, an unrestricted recovery or mutual information bound must account for that gap explicitly.

Corollary 1 (MI bound for a Bayes-complete local family). *Let X be uniform on K classes, and suppose the assumptions of Theorem 1 hold. For a radius vector ε in its grid, assume directly that $P_{\mathrm{guess}}^{\mathcal{H}_{\mathrm{loc}}(\varepsilon)}(X \mid Y) = P_{\mathrm{guess}}(X \mid Y)$. We define the clipped recovery advantage as $\varepsilon^{\star}_{\mathrm{loc}}(\varepsilon) = \max\{0, p^{+}_{\mathrm{loc}}(\varepsilon) - 1/K\}$. The mutual information then satisfies the following upper bound with confidence at least $1-\delta$.*

$$I(X;Y) \le \log_2(1 + K\varepsilon^{\star}_{\mathrm{loc}}(\varepsilon))\,. \tag{84}$$

Without this completeness assumption, $p^{+}_{\mathrm{loc}}(\varepsilon)$ remains an upper bound on the local family's recovery probability. Interpreting the converted value as a mutual information upper bound requires the additional assumption.

Proof of Corollary 1 On the confidence event from Theorem 1, the local-family recovery probability satisfies

$$\begin{aligned} P_{\mathrm{guess}}^{\mathcal{H}_{\mathrm{loc}}(\varepsilon)}(X \mid Y) &\le p^{+}_{\mathrm{loc}}(\varepsilon) \\ &\le 1/K + \varepsilon^{\star}_{\mathrm{loc}}(\varepsilon). \end{aligned} \tag{85}$$

The second relation is written as an inequality because the reporting convention clips negative advantage at zero. Since $p^{+}_{\mathrm{loc}} \le 1$, the clipped advantage is at most $1 - 1/K$. Bayes-completeness identifies the left side with unrestricted guessing, and Equation (74) then proves the corollary on the same event. The argument also applies simultaneously to any grid families known to satisfy the completeness assumption.

Learning bounds and certification bounds A learning analysis examines how a finite set of profiling traces affects the trained model. Certification conditions on the resulting model and uses independent attack traces to bound its population recovery probability. These analyses therefore address different sources of uncertainty. For a local family, the relaxed-margin event introduces additional conservatism because its probability can exceed the family's supremum recovery probability. More certification traces can improve statistical precision for this event probability. They need not remove the gap introduced by the relaxation or establish that training attained the best recovery probability in the family.

Example: What share separation establishes Consider a noiseless example with independent uniform byte shares S_1, S_2 and secret $X = S_1 \oplus S_2$. Each trace has one measurement channel and trace dimension $d = 700$, with the shares exposed at temporal positions 50 and 250. All other trace values are fixed. Either share alone is independent of X, while the pair determines X exactly. The two positions differ by 200 indices, so a contiguous patch must have temporal window length at least 201 to include both.

A predictor restricted to one fixed patch with temporal window length 100 observes at most one share, so its population recovery probability is 1/256. A finite-sample upper confidence bound can still exceed this value because of

statistical uncertainty. The same recovery bound requires a separate argument for a pooled family because aggregation can combine information from several patches. A patch containing both shares permits their joint use, but Bayes-completeness of a shared additive model still requires the stated representation and approximation assumptions. Repeated computation alone does not establish these assumptions.

To illustrate the reporting convention separately, a hypothetical certified advantage of 0.15 gives $\log_2(1 + 256 \cdot 0.15) \approx 5.300$ bits. An empirical advantage of 0.12 gives approximately 4.987 bits, leaving a difference of approximately 0.313 bits. These illustrative values are independent of the preceding masking example. Their difference reflects the statistical uncertainty and any family relaxation included in the chosen certificate. It does not directly measure the gap between the family's recovery probability and the Bayes optimum.

C.7 Composition across Targets and Traces

Combining evidence changes the recovery task because an attack can use several model outputs or traces to infer the same key. A certificate for one model's predicted class need not bound all information in its score vector. Similarly, separate guarantees for individual targets require additional assumptions before they can bound recovery from joint evidence. We first prove the post-processing inequality and then state the assumptions needed to combine recovery or mutual information bounds.

Composition across targets Let K denote the master key, $T_1, \ldots, T_\nu$ the target variables, and $Z_1, \ldots, Z_\nu$ their corresponding model output vectors, which provide soft evidence about the targets. We write $Z = (Z_1, \ldots, Z_\nu)$ and $B = \Phi(Z)$ for the output of a deterministic post-processing procedure, such as belief propagation with a fixed schedule. Public inputs are fixed or included in the conditioning. The symbol K continues to denote the target alphabet size, while ν counts the target variables.

Lemma 5 (Post-processing inequality for guessing probability). *For any deterministic post-processing output $B = \Phi(Z)$, the unrestricted key-recovery probability satisfies*

$$P_{\text{guess}}(\mathsf{K} \mid B) \leq P_{\text{guess}}(\mathsf{K} \mid Z). \tag{86}$$

Proof. Any guessing rule based on B can be applied to Z by first computing $\Phi(Z)$. It then has the same success probability. Taking the supremum over rules using B proves the inequality.

The lemma applies without independence among the evidence components. Its right-hand side is the unrestricted guessing probability from their joint values. Applying separate target certificates to this quantity requires a further composition argument.

Lemma 6 (Divide-and-conquer composition). *Suppose* K *is represented by the tuple of independent uniform targets* $(T_1, \ldots, T_\nu)$*, or by a known bijection of that tuple. Assume that the evidence vectors are mutually independent and the posterior factorizes as*

$$\Pr[\mathsf{K} = (t_1, \ldots, t_\nu) \mid Z] = \prod_{i=1}^{\nu} \Pr[T_i = t_i \mid Z_i]. \tag{87}$$

Suppose valid upper bounds $q_i \in [1/|\mathcal{T}_i|, 1]$ *on unrestricted guessing,* $P_{\text{guess}}(T_i \mid Z_i) \le q_i$*, hold with confidence at least* $1 - \delta_i$*. On their joint event, every deterministic post-processing satisfies*

$$P_{\text{guess}}(\mathsf{K} \mid B) \le \prod_{i=1}^{\nu} q_i. \tag{88}$$

Consequently, the conditional min-entropy defined through guessing probability satisfies the lower bound

$$\widetilde{H}_\infty(\mathsf{K} \mid B) := -\log_2 P_{\text{guess}}(\mathsf{K} \mid B) \ge \sum_i -\log_2 q_i. \tag{89}$$

The joint confidence is at least $1 - \sum_i \delta_i$*.*

Proof. A union bound gives the joint confidence event. On that event, posterior factorization and independence of the evidence give

$$\begin{aligned} P_{\text{guess}}(\mathsf{K} \mid Z) &= \mathbb{E}_Z \prod_i \max_{t_i} \Pr[T_i = t_i \mid Z_i] \\ &= \prod_i \mathbb{E}_{Z_i} \max_{t_i} \Pr[T_i = t_i \mid Z_i] \\ &= \prod_i P_{\text{guess}}(T_i \mid Z_i) \le \prod_i q_i. \end{aligned} \tag{90}$$

Lemma 5 extends this recovery upper bound to B. Applying the decreasing function $-\log_2$ then gives the stated lower bound on conditional min-entropy.

A uniform-target, Bayes-complete certificate for recovery from Z_i can provide $q_i = 1/|\mathcal{T}_i| + \bar{\varepsilon}_i^\star$. A certificate for the original model's top prediction alone need not do so, because another rule can use all entries of Z_i. The entropy terms above describe remaining uncertainty and are not the per-target BI values.

Lemma 7 (Information-based SASCA composition). *Let* K *be uniform on* $\mathcal{K}$*, with each* T_i *a deterministic function of the key and fixed public inputs. Suppose the evidence components are conditionally independent given the key and that each component's conditional distribution depends on the key only through its target. These assumptions give the factorization*

$$P_{Z|\mathsf{K}}(z \mid k) = \prod_i P_{Z_i|T_i}(z_i \mid T_i(k)). \tag{91}$$

Assume that valid mutual information upper bounds $I(T_i; Z_i) \le \eta_i$ in bits each hold with confidence at least $1 - \delta_i$. For any fixed belief-propagation procedure $B = \Phi(Z)$, the following recovery upper bound holds with confidence at least $1 - \sum_i \delta_i$.

$$P_{\text{guess}}(\mathsf{K} \mid B) \le \min\left\{1, \frac{1}{|\mathcal{K}|} + \sqrt{\frac{\ln 2}{2} \sum_i \eta_i}\right\}. \tag{92}$$

Without conditional independence, the post-processing inequality still holds, but this sum-based bound need not follow.

Proof. Work on the joint event where the information bounds hold. Conditional independence and the chain rule give

$$\begin{aligned} I(\mathsf{K}; Z) &= \sum_i I(\mathsf{K}; Z_i \mid Z_{<i}) \\ &\le \sum_i I(\mathsf{K}; Z_i) = \sum_i I(T_i; Z_i) \le \sum_i \eta_i. \end{aligned} \tag{93}$$

The inequality uses $I(\mathsf{K}; Z_i \mid Z_{<i}) = I(\mathsf{K}; Z_i) - I(Z_i; Z_{<i})$ under conditional independence. The next equality follows because the conditional distribution of Z_i depends on the key only through T_i.

For any posterior on a finite uniform key space, its maximum probability exceeds $1/|\mathcal{K}|$ by at most its total variation distance from the uniform prior. Pinsker's inequality and Jensen's inequality therefore give

$$P_{\text{guess}}(\mathsf{K} \mid Z) \le \frac{1}{|\mathcal{K}|} + \sqrt{\frac{\ln 2}{2} I(\mathsf{K}; Z)}. \tag{94}$$

Substituting the information upper bound and using the fact that recovery probability is at most one gives the claimed bound for Z. Lemma 5 then gives the same upper bound for B.

For uniform T_i, a Bayes-complete certificate for the full evidence Z_i supplies $\eta_i = \log_2(1+|\mathcal{T}_i|\bar{\varepsilon}_i^\star)$ through Lemma 1. Other justified information bounds can be used instead. Stating the η_i explicitly avoids assuming that ordinary per-model accuracy already certifies the information in the soft evidence.

Composition across traces Several traces can reveal more about the same secret than any individual trace. Under conditional independence, valid single-trace mutual information bounds can be combined to give an upper bound on recovery from multiple traces. This bound applies under the stated distributional assumptions and need not equal the recovery probability of a particular attack.

Proposition 5 (Multi-trace recovery bound). *Let X be uniform on K values and let $Y^{(1)}, \ldots, Y^{(n_a)}$ be n_a attack traces that are conditionally independent given X. Suppose that, on a joint confidence event, $I(X; Y^{(j)}) \le \eta_j$ for each*

trace distribution. For any family $\mathcal{H}$ of rules using the complete trace tuple, let $\mathrm{SR}^{\mathcal{H}}(n_a)$ denote its supremum recovery probability. This probability satisfies

$$\mathrm{SR}^{\mathcal{H}}(n_a) \leq \min\left\{1, \frac{1}{K} + \sqrt{\frac{\ln 2}{2}\sum_{j=1}^{n_a}\eta_j}\right\}. \tag{95}$$

If the traces also have identical conditional distributions, a common information upper bound $\eta = \log_2(1 + K\bar{\varepsilon}^\star)$ obtained under the assumptions of Corollary 1 gives

$$\mathrm{SR}^{\mathcal{H}}(n_a) \leq \min\left\{1, \frac{1}{K} + \sqrt{\frac{n_a \ln 2}{2}\log_2(1 + K\bar{\varepsilon}^\star)}\right\}. \tag{96}$$

The common information bound retains its original confidence event because the same population bound applies to every trace distribution. If the η_j are certified separately with failure probabilities δ_j, a union bound gives joint confidence at least $1 - \sum_j \delta_j$.

Proof. The same conditional-independence argument used above yields

$$I(X; Y^{(1:n_a)}) \leq \sum_{j=1}^{n_a} I(X; Y^{(j)}) \leq \sum_j \eta_j. \tag{97}$$

Let u_K denote the uniform prior and let TV denote total variation distance. The unrestricted guessing advantage satisfies

$$\begin{aligned} P_{\mathrm{guess}}(X \mid Y^{(1:n_a)}) - 1/K &\leq \mathbb{E}\,\mathrm{TV}(P_{X|Y^{(1:n_a)}}, u_K) \\ &\leq \sqrt{\frac{\ln 2}{2} I(X; Y^{(1:n_a)})}. \end{aligned} \tag{98}$$

The second inequality follows by applying Pinsker's inequality to each posterior and Jensen's inequality to the expectation [14]. Recovery within any specified family is at most the unrestricted guessing probability. Substituting the information upper bound therefore proves both forms on the same confidence event as the original information bounds.

For a common positive information upper bound η and a desired recovery probability $s > 1/K$, rearranging the recovery inequality gives the necessary lower bound

$$n_a \geq \frac{2(s - 1/K)^2}{\ln 2\, \eta}. \tag{99}$$

The right-hand side must be rounded upward because the number of attack traces is an integer. This lower bound gives a necessary number of traces for attaining s, but does not establish that any attack attains s with that many traces. If $\eta = 0$, conditional independence keeps unrestricted recovery at the random-guessing probability for every finite number of traces. When $K\bar{\varepsilon}^\star \ll 1$, the denominator $\ln 2\, \eta$ is approximately $K\bar{\varepsilon}^\star$.

C.8 Correlated Evidence and Complete-Pipeline Certification

Overlapping leakage, shared computation, and masking can introduce dependencies that prevent product composition. Pairwise covariance can reveal some dependencies, but zero covariance does not establish independence or posterior factorization. The post-processing inequality remains valid in this setting. Combining marginal information bounds by summation still requires the conditional-independence assumptions in Lemma 7.

A different approach certifies the complete recovery procedure directly, using one success or failure per campaign. It avoids intermediate composition assumptions, while requiring independent campaigns for the fixed pipeline being evaluated.

Proof of the complete-pipeline certificate

Proof (Proof of Proposition 3). We treat each complete attack campaign as one statistical trial. We condition on all data used to train, select, and tune the pipeline $\mathcal{P}$, including its models, score combination, belief-propagation schedule, and enumeration budget. The pipeline is then fixed. By assumption, the campaigns $\mathcal{A}_1, \ldots, \mathcal{A}_{n_c}$ are i.i.d. and independent of these development data. We define the success indicator for campaign i as

$$Z_i := \mathbf{1}\{\mathcal{P}(\mathcal{A}_i) \text{ recovers the key}\}. \tag{100}$$

These indicators are i.i.d. Bernoulli variables with mean p_{pipe}. The singleton argument in Appendix B.3 therefore gives the following upper bound with probability at least $1 - \delta$.

$$p_{\text{pipe}} \leq U_{\text{binom}}\left(\sum_{i=1}^{n_c} Z_i, n_c, \delta\right). \tag{101}$$

Averaging over the development data preserves the confidence and proves the proposition.

Direct certification avoids the relaxations introduced by intermediate composition bounds. Its statistical uncertainty can nevertheless be large when few independent campaigns are available. To obtain a one-sided Hoeffding margin of at most $a > 0$, it suffices to use $n_c \geq \ln(1/\delta)/(2a^2)$ campaigns. Here a is a statistical margin in recovery probability, while the score radius controls model deviations. Several pipelines fixed independently of the certification campaigns may share the same campaign set if the confidence allocation covers them simultaneously. A pipeline changed using certification outcomes requires a suitable adaptive method or new independent campaigns.

C.9 Four-Byte Recovery on Raw ASCAD Traces

The raw-ASCAD experiment evaluates an attack that combines evidence across bytes. It starts from raw traces with trace dimension $d = 100{,}000$. A points-of-interest scan examines all sixteen bytes over these traces, and compact two-share

Table 10: The table reports empirical four-byte recovery on raw ASCAD traces using fixed two-share models for bytes 0, 1, 5, and 15. Each row summarizes 250 evaluated campaigns with the specified number of traces per campaign. Rank zero denotes the first candidate. The reported values describe empirical performance without asserting a population confidence bound.

	Four-byte composition			
Traces	Successes	Empirical rate	Median rank	Mean rank
100	40/250	0.160	7.0	23.43
200	94/250	0.376	1.0	5.37
500	191/250	0.764	0.0	0.37
1,000	206/250	0.824	0.0	0.18
2,000	221/250	0.884	0.0	0.12

models then use selected windows around the unmasked S-box, masked S-box, and mask locations. Thus, the source trace dimension describes the data available for window selection, while each model's input dimension is determined by its selected windows. Validation selects bytes 0, 1, 5, and 15 for the four-byte attack.

The attack combines the selected models' outputs through a product score and enumerates four-byte candidates. Table 10 reports recovery in 191/250 campaigns at 500 traces, 206/250 at 1,000, and 221/250 at 2,000. The median candidate rank is zero from 500 traces onward, where rank zero denotes the first candidate.

These results show that the specified attack recovers several bytes from this trace source. Using a product score does not verify the independence or posterior-factorization assumptions of Lemma 6. Nor do the reported counts alone establish the independent-campaign protocol needed for a population certificate. The results remain empirical diagnostics and use different windows and models from the single-byte suite results in Section 5.

D Additional Experimental Details

This appendix provides additional details on the datasets, attacker suites, and evaluation protocols used in Section 5. It first describes the benchmark inputs and data separation, then examines how the bounds depend on attack-set size, temporal window length, and attacker-family composition. It also compares trained models with Bayes reference rules on synthetic data and with additional attackers on measured traces. The final two subsections describe score-radius calibration and the EPI temperature experiment. Where these experiments use different models or attack traces from Table 7, we state the differences when interpreting their results.

D.1 Datasets, Data Separation, and Attacker Suites

Benchmark traces. Each benchmark in Table 7 targets exact recovery of an 8-bit secret intermediate, so the secret alphabet has size $K = 256$. Following

Sections 2.1 and 5.1, the trace dimension d is the number of values in each trace supplied to the model. Preprocessing can therefore make this dimension differ from the number of values in the original recorded trace. The attack-set size instead denotes the number of traces used for evaluation. The benchmarks use the following trace dimensions.

- **ASCAD fixed key** [5] provides masked AES traces with controlled desynchronization levels of 0, 50, and 100. Each variant is evaluated at trace dimension 700.
- **ASCAD random key** [5] provides variable-key AES traces. The main EstraNet suite evaluates these traces at dimension 1,400.
- **AES-HD** [26] provides traces from an FPGA implementation of AES. The evaluated trace dimension is 1,250.
- **AES-RD** [13] provides AES traces with random delays. The evaluated trace dimension is 3,500.
- **DPAv4** [7] provides traces from a smart-card implementation of AES. We evaluate these traces at dimension 4,000 using the benchmark's procedure for assigning target labels.
- **CHES-CTF-2025** [6] provides challenge traces. The main table uses trace dimension 3,000, while a separate experiment increases the trace dimension to 7,000.

The CHES preprocessing uses the mean of the attack traces. We therefore report its computed lower and upper bounds as numerical diagnostics. A PAC guarantee additionally requires establishing independence between preprocessing and the traces used for certification.

Data separation. Profiling traces are used for training, and validation traces are used for model comparison and selection. The evaluated suite and its preprocessing must be fixed independently of the attack traces used for certification. Under this separation, repeating the same confidence calculation for the fixed suite leaves its guarantee unchanged. If certification traces influence model selection or adaptation, increasing the reported suite size afterward does not generally restore the required independence. Appendix B.1 states the corresponding evaluation conditions. The score-radius experiment uses an additional calibration partition, as described in Appendix D.6.

Trained suites. The six benchmarks comprising ASCAD d0, d50, and d100, AES-HD, AES-RD, and DPAv4 each use a `TCHES20-M80` suite containing 40 MLP and 40 CNN models. The `ASCADr-E11` and `CHES25-E11` suites each contain 11 trained EstraNet models and use 100,000 attack traces. These two suites are evaluated on their respective datasets, with the CHES results retaining the preprocessing qualification above. The Transformers used in the selected-model local-family experiment are separate from the M80 suites. Consequently, comparisons across benchmarks depend on both the measured traces and the trained models included in each evaluation.

Baseline calculations at reduced trace dimensions. The additional ePI, eHI, MLP-PI, and GKOV MI calculations use traces reduced to dimensions $d' = 10$, 50, and 100. These calculations provide numerical comparisons with BI, although they measure different quantities and use different input representations from the main BI evaluation. For example, the reported GKOV estimates on ASCAD random key decrease from 2.850 to 2.017 to 1.196 bits at these three dimensions, respectively. A finite or positive estimate alone does not establish that an estimator has been implemented correctly. Attributing the differences solely to trace dimension also requires matching the underlying data and the procedures used for projection.

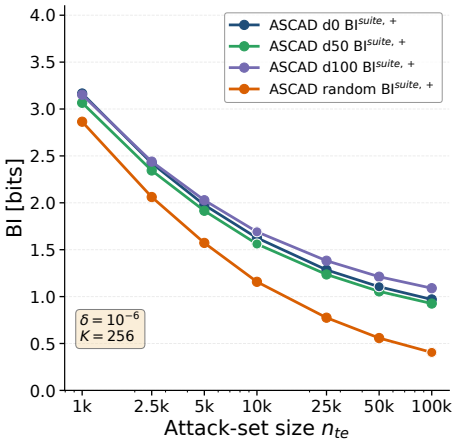


Fig. 4: The curves show BI upper bounds for four ASCAD variants as the assumed attack-set size increases at fixed empirical recovery probabilities and suite sizes.

D.2 Effect of Attack-Set Size

Figure 4 examines how attack-set size affects the BI upper bound for the four ASCAD variants when the empirical recovery probability and suite size remain fixed. For each curve, the assumed attack-set size ranges from 1,000 to 100,000, with $\delta = 10^{-6}$. The empirical recovery probabilities and suite sizes used here differ from those in Table 7. The curves therefore illustrate the effect of statistical concentration under these separate evaluation settings.

The upper bounds decrease as the assumed attack-set size grows because the statistical margin decreases. Their relative positions depend on the fixed inputs used for each curve, so their ordering does not rank the results in the main table. This calculation illustrates the reduction in uncertainty at a fixed empirical recovery probability. When new attack traces are collected, the empirical recovery probability can also change.

D.3 Comparing the Benchmark Results

In Table 7, AES-RD has a maximum empirical recovery probability of 0.13968, an empirical BI value of 5.160 bits, and a BI upper bound of 5.351 bits. The statistical margin expressed in bits is therefore approximately 0.191. The BI upper bounds for ASCAD d50, d100, and random key are 2.123, 2.134, and 2.313 bits, respectively. These values describe the evaluated attacker families at their corresponding attack-set sizes. Their ordering alone does not establish which physical countermeasure caused the differences.

DPAv4 illustrates the effect of statistical uncertainty with a small attack set. Its 500 attack traces yield a BI upper bound of 4.549 bits and a margin ratio of $R_{\text{margin}} = 3.36$. Thus, the statistical margin exceeds the empirical recovery advantage. The value of 4.549 bits includes this uncertainty and expresses an upper bound on recovery probability through the conversion in Section 5.2.

Figure 5 compares BI with the estimates computed at reduced trace dimension under a separate set of evaluation settings. Since these metrics measure different quantities, their ordering need not agree with the BI ordering. On AES-RD, the four displayed baseline estimates are negative, while the suite's empirical recovery probability exceeds the random-guessing baseline. Mutual information is nonnegative, so a negative MI estimate reflects estimation error. This property does not make the other metrics interchangeable with BI. The figure consequently provides a descriptive comparison, and assessing estimator accuracy or ranking device security would require matched evaluation conditions.

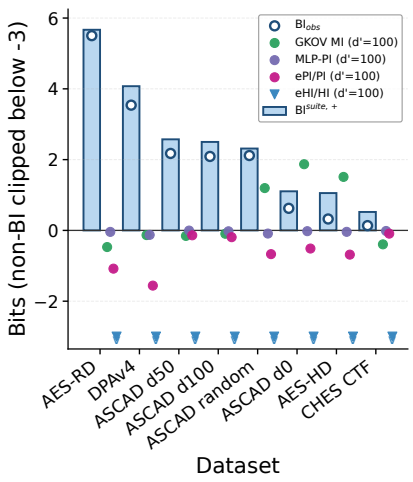


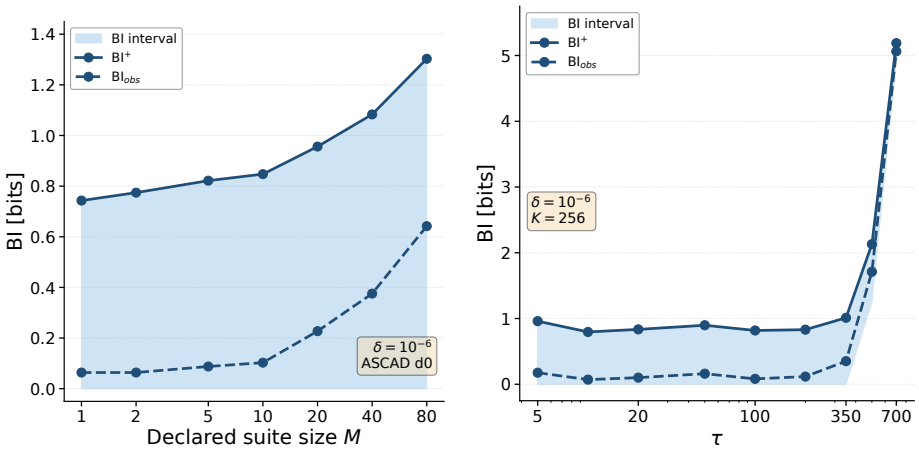
Fig. 5: The figure compares empirical BI values and finite-suite BI upper bounds with estimates computed at reduced trace dimension $d' = 100$. The plotted BI values use different evaluation settings from Table 7. Values below the lower display limit are clipped, and the CHES upper bound remains a numerical diagnostic.

D.4 Sensitivity to the Evaluated Family

Statistical margin. When the empirical recovery probability exceeds the random-guessing baseline, the ratio

$$R_{\text{margin}} = \frac{p^{+}_{\text{suite}} - p_{\text{obs}}}{p_{\text{obs}} - 1/K} \tag{102}$$

compares the statistical margin with the empirical recovery advantage on the probability scale. In the main table, this ratio is below one for AES-RD and ASCAD random key and above one for the other six rows, including the numerical CHES result. When empirical recovery is at or below the random-guessing baseline, we report the interval and margin directly, following Section 3.4. Increasing the attack-set size can reduce statistical uncertainty for a fixed family. Expanding the family can instead increase the upper bound because it can introduce more successful models and change the confidence allocation.



(A) Suite size (B) Window position

Fig. 6: **(A)** Empirical BI and upper bounds for the first M models in a fixed ASCAD d0 suite evaluated on 25,000 attack traces. As M increases from one to 80, empirical BI increases from 0.064 to 0.642 bits, while the upper bound increases from 0.743 to 1.303 bits. **(B)** Mean empirical BI and recomputed lower and upper bounds across five random seeds in a separate ASCAD d0 CNN experiment. Each run evaluates a window position selected using profiling traces.

Temporal-window diagnostic. The temporal-window experiment in Figure 6B uses a separate pretrained CNN pipeline on ASCAD d0. It evaluates five random seeds with 25,000 attack traces per seed. For each window length τ, the window position is selected using profiling traces. Here τ denotes the number of consecutive sampling positions within the window, consistent with Section 4.1. The plotting procedure computes KL-binomial lower and upper bounds for the model evaluated at the selected position, with a multiplicity factor equal to the number of positions tested. It then averages the empirical BI values and the computed bounds across seeds. Thus, each run evaluates its selected position rather than maximizing success across all positions on the attack set.

The mean BI lower bound remains zero through $\tau = 350$, while the mean BI upper bound increases to approximately 2.131 bits at $\tau = 500$ and 5.188 bits at $\tau = 700$. A zero BI lower bound indicates that this calculation does not establish population recovery above the random-guessing baseline. It leaves open whether an attack using such a window can achieve greater recovery success. The increase at larger windows shows the sensitivity of the computed bounds to window length in this pipeline. It does not establish a general minimum window length for exploiting leakage. These models and attack traces also differ from the ASCAD d0 evaluation using 10,000 traces in the main table.

Suite-size diagnostic. Figure 6A examines how the BI upper bound changes as models are added to an ASCAD d0 suite evaluated on 25,000 attack traces.

The models have a fixed order, and each evaluated subset contains the first M models for $M \in \{1, 2, 5, 10, 20, 40, 80\}$. For each subset, the figure reports the empirical BI value and the KL-binomial BI upper bound. This experiment uses a different attack-set size from the main-table evaluation with 10,000 traces. The model order must be fixed independently of the certification traces for the stated confidence calculation to apply.

The empirical maximum remains unchanged across some successive suite sizes even though the upper bound increases. These segments show the additional statistical cost of allocating confidence across more models. Other increases also reflect newly included models with higher empirical recovery probabilities. The upper bound for all 80 models is 1.303 bits in this evaluation using 25,000 traces. Thus, expanding the suite can reveal stronger attacks while increasing the statistical margin, and both effects contribute to changes in the bound.

Architecture diagnostic. Figure 7 compares trained attacker families that progressively include linear models, MLPs, and broader CNN suites in a separate evaluation. On ASCAD d0, empirical BI increases when the family expands from linear models to include MLPs and then remains near one bit. On AES-RD, empirical BI increases more gradually across the evaluated families. These results describe the separately evaluated models and attack traces, so they cannot be substituted for the M80 results in the main table.

A constant empirical maximum shows that the added models did not improve recovery on the evaluated attack traces. A model outside these families may still achieve greater success. Relating these values directly to the main table requires matching the training and evaluation protocols.

D.5 Comparisons with Bayes Reference Rules and Additional Attackers

Synthetic data generated from a known distribution allow us to evaluate an explicit Bayes reference rule. The corresponding true posterior is unavailable for the measured-trace benchmarks, so we instead compare their trained suites with additional attackers, which we call challengers. These experiments assess the evaluated models' performance relative to specific alternatives. Establishing Bayes-completeness additionally requires equality between the family's supremum population recovery probability and unrestricted Bayes vulnerability, as defined in Section 3.5.

In the Gaussian byte experiment, the reference value is the empirical recovery probability of the Bayes rule for the known distribution, and the comparison model is a learned Gaussian template. Their mean difference across three random seeds is approximately $7.3 \cdot 10^{-5}$. In the masked/locality experiment, the reference value is the average of the maximum known posterior probability over the evaluated traces. The trained predictor using a window length of 65 has a mean empirical recovery probability of 0.925720, which differs from the reference by approximately $6.8 \cdot 10^{-4}$. These results show close numerical agreement in the evaluated settings. The formal Bayes-completeness results depend on the

Table 11: The table compares Bayes reference values with trained-model recovery probabilities on synthetic data, averaged over the runs in each experiment. The difference is computed from the unrounded reference and model values.

Setting	K	Bayes reference	Trained model	Difference
Gaussian byte	256	0.989577	0.989503	$7.3 \cdot 10^{-5}$
Masked/locality	2	0.926401	0.925720	$6.8 \cdot 10^{-4}$

Table 12: The table compares suite and challenger empirical recovery probabilities under evaluation settings that differ from Table 7. The last column indicates whether the maximum empirical challenger recovery probability exceeds the listed suite upper bound.

Dataset	Suite empirical	Upper bound	Challenger	Exceeds bound?
AES-RD	0.17710	0.19873	0.00600	No
ASCAD d0	0.00610	0.00965	0.00900	No
ASCAD random	0.01693	0.01941	0.01693	No
CHES-CTF-2025	0.00429	0.00561	0.00438	No

assumptions and proofs in Appendix C.4. The masked experiment also evaluates the oracle prediction rule itself.

Across runs, the mean upper bound on its recovery probability is 0.939762, and the mean BI upper bound is 0.910367 bits. These are averages of the oracle's per-run confidence bounds on their respective scales. They therefore describe uncertainty in the oracle evaluation and do not bound the trained predictor's approximation gap.

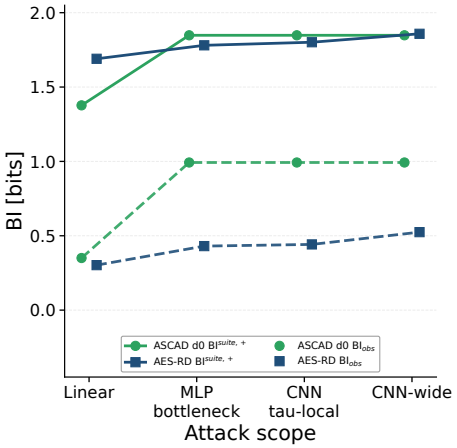


Fig. 7: The curves show empirical BI values and upper bounds as the evaluated attacker families expand to include additional architectures. A constant empirical maximum describes only the tested models and does not establish Bayes-completeness or extend the certificate to unevaluated models.

Table 12 summarizes 35 challenger comparisons. The AES-RD and ASCAD d0 suite settings differ from those in Table 7, and some suites and challengers use different attack sets. Each listed upper bound therefore applies to the suite evaluation used in that comparison. The ASCAD random-key comparisons include reference models from the EstraNet suite, so some comparisons reuse models already included in the suite.

Every empirical challenger recovery probability in these comparisons is below the corresponding listed upper bound. However, the ASCAD d0 and

Table 13: The table reports calibration and certification trace counts for the selected-model score-radius experiment. Each count applies separately to every model listed in the corresponding row.

Benchmark	Models	Calibration	Certification
ASCAD d50	MLP, CNN, Transformer	2,000	8,000
ASCAD d100	MLP, CNN, Transformer	2,000	8,000
ASCAD random	MLP, CNN, Transformer	20,000	80,000
AES-RD	MLP, CNN	2,500	10,000
AES-RD	Transformer	5,000	20,000

CHES challenger estimates exceed the corresponding empirical suite maxima. The absence of a challenger estimate above the suite upper bound therefore does not imply that all tested challengers have lower empirical recovery probabilities. These comparisons do not establish saturation or Bayes-completeness. The fixed-suite guarantee also does not extend to a challenger solely because its empirical recovery probability lies below the suite's upper bound.

D.6 Score-Radius Calibration

The selected-model experiment uses twelve reference models, comprising one MLP, one CNN, and one Transformer for each of ASCAD d50, ASCAD d100, ASCAD random key, and AES-RD. For each model, the reserved traces are divided into a calibration partition containing 20% of the traces and a certification partition containing the remaining 80%. Table 13 reports the trace counts for these partitions. These counts differ from those used in the full M80 local-family experiment in Section 5.6.

Before certification, the radius grid is fixed as

$$\mathcal{E} = \left\{0, 10^{-5}, 3 \cdot 10^{-5}, 10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}\right\}. \tag{103}$$

For each reference model and radius, the upper bound uses failure probability $\delta/(12|\mathcal{E}|) = 10^{-6}/108$. A union bound therefore gives simultaneous coverage across all twelve reference-model curves and nine radii, provided the models, preprocessing, score normalization, and radius grid satisfy the independence conditions in Theorem 1. This joint confidence allocation covers the separate dataset-specific guarantees. Each model neighborhood remains associated with its own evaluation distribution.

Two criteria determine the reported radius. First, the relaxed-margin count must increase relative to radius zero for at least 75% of the reference models. Second, the radius must be at least as large as the maximum normalized score difference observed when inference is repeated with different batch sizes. The first criterion holds at 10^{-4}. The maximum observed score difference is approximately $1.905 \cdot 10^{-4}$, so $3 \cdot 10^{-4}$ is the smallest grid value satisfying both criteria. We therefore select $\varepsilon_{\text{cal}} = 3 \cdot 10^{-4}$ and compute the upper bounds using the separate certification partitions.

Calibration selects a score radius based on the evaluated reference models and implementation variations. To establish that another model belongs to a

Table 14: The table reports how many reference models have a larger relaxed-margin count at each radius than at radius zero on calibration traces.

Radius ε	Models	Fraction
10^{-5}	4/12	33.3%
$3 \cdot 10^{-5}$	7/12	58.3%
10^{-4}	9/12	75.0%
$3 \cdot 10^{-4}$	10/12	83.3%
10^{-3}	11/12	91.7%
$3 \cdot 10^{-3}$	12/12	100%
10^{-2}	12/12	100%
$3 \cdot 10^{-2}$	12/12	100%

Table 15: The table reports the variation in raw PI and recalibrated EPI across the temperature sweep, measured in bits. Every predicted label remains unchanged, so the empirical BI value is constant on each benchmark.

Benchmark	Raw PI variation	EPI variation
ASCAD d0	4.23	$3.05 \cdot 10^{-10}$
ASCAD d50	23.07	$2.51 \cdot 10^{-10}$
ASCAD d100	22.48	$9.43 \cdot 10^{-11}$
AES-HD	5.30	$1.78 \cdot 10^{-10}$
AES-RD	27.31	$7.02 \cdot 10^{-11}$
DPAv4	20.71	$8.57 \cdot 10^{-10}$

certified neighborhood, its normalized class scores must satisfy the uniform score-distance condition in Definition 5 over every admissible trace. Appendix C.3 gives sufficient deterministic bounds for establishing this membership. As discussed in Section 5.6, the Transformer models use logits divided by a fixed constant determined from training traces, while the MLP and CNN models use normalized probability scores. Equal numerical radii therefore need not represent equivalent changes in model behavior across architectures.

D.7 Comparison with Effective PI

For each of the six TCHES20 benchmarks, we keep the evaluated model and attack traces fixed while varying the temperature from 0.05 to 20. At each temperature, we recompute raw PI and recalibrate EPI using the probability-fitting procedure in Section 5. This fitting uses profiling traces that are separate from the attack traces. Table 15 reports the difference between the largest and smallest metric values across the temperature sweep.

Raw PI changes substantially across the temperature sweep, while EPI remains nearly constant after recalibration. Empirical BI remains unchanged because positive temperature scaling preserves every predicted label and hence every exact-recovery outcome. The finite-suite confidence bounds also remain unchanged when the suite, attack-set size, and confidence allocation are fixed. A local score family additionally depends on numerical score differences, which can change under temperature scaling. Its radius must therefore be interpreted together with the specified score representation and normalization.