

Generalized Power Attacks against Crypto Hardware using Long-Range Deep Learning

Elie Bursztein¹, Luca Invernizzi², Karel Král², Daniel Moghimi¹,
Jean-Michel Picod² and Marina Zhang¹

¹ Google, Sunnyvale, USA, scaaml@google.com

² Google, Zurich, Switzerland, scaaml@google.com

Abstract. To make cryptographic processors more resilient against side-channel attacks, engineers have developed various countermeasures. However, the effectiveness of these countermeasures is often uncertain, as it depends on the complex interplay between software and hardware. Assessing a countermeasure’s effectiveness using profiling techniques or machine learning so far requires significant expertise and effort to be adapted to new targets which makes those assessments expensive. We argue that including cost-effective automated attacks will help chip design teams to quickly evaluate their countermeasures during the development phase, paving the way to more secure chips.

In this paper, we lay the foundations toward such automated system by proposing GPAM, the first deep-learning system for power side-channel analysis that generalizes across multiple cryptographic algorithms, implementations, and side-channel countermeasures without the need for manual tuning or trace preprocessing. We demonstrate GPAM’s capability by successfully attacking four hardened hardware-accelerated elliptic-curve digital-signature implementations. We showcase GPAM’s ability to generalize across multiple algorithms by attacking a protected AES implementation and achieving comparable performance to state-of-the-art attacks, but without manual trace curation and within a limited budget. We release our data and models as an open-source contribution to allow the community to independently replicate our results and build on them.

Keywords: Deep Learning · Side-Channel Analysis · AES · ECC

1 Introduction

Cryptographic co-processors, which are widely used to perform security-sensitive operations, can be vulnerable to side-channel attacks. These attacks aim to extract the secrets these chips safeguard, such as AES [KJJ99] or RSA keys [Koc96]. Broadly speaking, side-channel attacks recover secret data by collecting signals such as timing, power consumption [MOP08], and electromagnetic emissions [QS01] while the chip runs computations. These signals are later processed using statistical methods [CRR03] or machine learning techniques [MPP16] to recover the targeted secret information (e.g., the AES key). Attackers can then use the recovered secrets to bypass vital security features such as secure boot,¹ remote attestation [Fid15, MSEH20], and identity protection.²

In recent years, *side-channel attacks assisted with machine learning (SCAAMLs)* [PB19] due to their superior accuracy, reduced manual work, and lesser need of domain knowledge, have started to replace profiling attacks such as Template attacks [CK14, CRR03]. In particular, SCAAML attacks have proved to be effective at recovering AES keys, even

¹<https://source.android.com/docs/security/features/verifiedboot>

²<https://source.android.com/docs/security/features/biometric/measure>

when facing strong masking countermeasures [ERR⁺18, MPP16, CDP17].

Despite their clear effectiveness against specific targets, several obstacles limit SCAAMLs broad adoption during the product development cycle [PPM⁺23], where engineers need to be notified of potential leaks in a matter of hours or days to stick to the production schedule, including:

- **Lack of cross-algorithm generality:** SCAAMLs have not been reported to generalize beyond a single protected cryptographic algorithm [LBM15, MPP16, CDP17, ERR⁺18, MS23, WHJ⁺21, HGG19, ZS20, WP20, PP21, WAGP20, AGF23, WPP22, ZBHV20, PCP20, RBA20, ZSX⁺20, BCH⁺20, LZC⁺21]. There is currently no known machine learning technique that is able to attack both highly protected AES and ECC for example.
- **Lack of cross-implementation generality:** Current attacks rely on custom ML architectures tailored to a very specific target [PP21, WAGP20, AGF23, WPP22, ZBHV20, PCP20]. Targeting a different implementation that uses different countermeasures requires to manually modify the attack.
- **High-expertise requirements:** So far SCAAMLs require expertise to not only modify the neural network architecture and its objectives but most of them also require expert manual pre-processing of the traces (e.g., [LZC⁺21]) so they can be used by the neural network.

In this paper, we provide the first step toward addressing these limitations by proposing a novel deep-learning architecture, GPAM (Generalized Power Analysis Model), that is able to perform fully automated power side-channel attacks against multiple protected algorithms, namely ECC and AES, countermeasures, and implementations. GPAM is designed to work on raw traces straight from the oscilloscope doing away with the expensive requirement of expertly preprocessing traces before performing attacks. Our novel architecture, presented in Section 5, combines *temporal patchification* [LMW⁺22] to process the very long traces generated by algorithms incorporating countermeasures, state-of-the-art Transformer encoder blocks [HDLL22] to efficiently identify long-range trace data relationships, and multi-task learning [Rud17] to allow the model to attack masked implementations.

We demonstrate GPAM effectiveness by carrying out power analysis attacks against four protected ECDSA implementations in Section 6. These implementations countermeasures range from a simpler-to-defeat constant-time countermeasure to masking protections that are considered resistant to side channels attacks [Cor99, PCBP21, GRV17, RLMI21]. Specifically, GPAM is able to recover the four most significant bits of the secret scalar with an accuracy between 71.86% to 96.39% depending on the targeted implementation. At that level of accuracy, combining the model predictions confidence with a lattice attacks is enough to recover the full secret key [HGS01, NS02], as demonstrated in Section 6.8. To the best of our knowledge, this is the first time that these highly-protected ECDSA implementations have been proven to be vulnerable to power side-channel attacks, demonstrating that GPAM architecture is not only general but also highly effective at attacking state-of-the-art hardware defenses. We note that generalized models, such as GPAM, fulfill a different need than custom attack models. Custom attack models excel at uncovering vulnerabilities in high-value targets, but require the expertise of side-channel specialists. Instead, generalized models empower non-experts, such as implementation engineers, to evaluate the side-channel security of their designs without specialized attack knowledge. As such, they *complement*, and not supersede, custom attack models.

In Section 7, we show-case GPAM’s ability to generalize to multiple cryptographic algorithms without architectural changes by successfully recovering masked AES keys. When compared with state-of-the-art attacks that rely on manual trace pre-processing and hand-tuning (which GPAM does not require), GPAM achieves comparable performance.

Last but not least in Section 6.8, we demonstrate that GPAM generalization capabilities

also extend beyond white-box attacks by demonstrating its ability to recover hardware masked ECC scalar in a black-box settings.

Overall, the sum of our experimental results highlights that this new generation of generalized automated attacks is competitive with algorithm-specific state-of-the-art approaches for evaluating power leakage countermeasures. Moreover as discussed in Section 5.3, the operational costs of adapting GPAM to a new target via automated hyper-tuning, a few hours of GPU time, is considerably lower than hiring side-channel experts. Our attack generality, speed, and cost-effectiveness move us closer to more secure chips by empowering design teams to incorporate automated countermeasure testing as part of the development process.

To allow the community to independently replicate our results and get us closer to the standardization of fully-automated side-channel leakage evaluations we open-source both our models and datasets under the Apache 2 Licence at [anonymized].

Ethics This research was intentionally performed on research implementations, not production ones. Accordingly, these results do not warrant a coordinated responsible disclosure.

2 Background

This section provides the key background information on cryptography, side-channel attacks, and deep learning needed to understand the paper.

2.1 Elliptic-curve cryptography

Elliptic-curve cryptography (ECC), which supports both key exchange and digital signatures, comprises public parameters, and a public/private key pair. Public parameters include an elliptic curve E , a point G on the curve, and the integer order n of G over E . The secret key d is a random integer satisfying $1 < d < n - 1$. The public key is calculated as $Q = d \times G$ (\times is the scalar multiplication operation supported by curve E). As relevant to this paper, to generate a signature for a message hash h , ECDSA algorithm chooses a random secret k such that $1 < k < n - 1$, computes $(x, y) = k \times G$, $r = x \bmod n$, and $s = k^{-1}(h + r \cdot d) \bmod n$, and outputs (r, s) as the signature pair.

It is critical for the private key d and the per-message random secret k to remain secret. An attacker who acquires one instance of k for a known signature can simply calculate the private key as $d = r^{-1}(s \cdot k - h) \bmod n$. An attacker who can recover part of k can apply lattice-based cryptanalysis to recover the private key from partial knowledge of k from several signature-generation operations [HGS01, NS02].

2.2 Side-Channel Attack and Defense

Side-channel attacks (SCA) target the execution of cryptographic algorithms [MOP08, Koc96]. During the execution, certain physical signals may be generated by intermediate computations that depend on the secret data bits being processed. The attacker can build a distinguisher that identifies which signals are related to which secret bits. This is typically accomplished by attacking each segment of the key separately. For instance, we build a distinguisher that can find the correct key byte for AES, which can be repeated for each of the 16 different key bytes of the 128-bit AES key.

There are two primary scenarios for constructing a distinguisher: In direct attacks,

such as SPA [MOP08] or DPA [KJJ99], the attacker attempts to retrieve the key from traces without prior modeling of the target. In profiling-based attacks, e.g., Template attacks [CRR03], the attacker first constructs a model based on previous observations of the target (or a similar one). We focus on white-box profiling-based attacks, which are valuable for assessing the security of an implementation against a strong, well-informed attacker.

One can use masking countermeasures to mitigate side-channel attacks by disrupting the statistical correlation between intermediate values and the physical signal, e.g., power consumption. To achieve this, implementations can generate a random value and combine it with secret parameters and intermediate values during computation. As a result, the computation is carried out using blinded secrets instead of cleartext ones.

A common protection for ECC implementations is to randomize the secret integer (d or k) during scalar multiplication [Cor99]. For this, implementations can add a random multiple of the curve order n to the private integer k as $k' \rightarrow k + r \cdot n$. Later on, when computing the scalar multiplication $k' \times G$ as in the signature generation, it results as $(k + r \cdot n) \times G = k \times G + r \cdot n \times G$. Since $n \times G$ is equal to the point at infinity (the identity element), the expression simplifies to $k \times G$. Randomizing the secret integer can also be achieved using the euclidean division $k = \lfloor k/r \rfloor \cdot r + (k \bmod r)$, or the secret integer can be divided into multiple random shares for extra security $k = k_1 + \dots + k_m$. In this paper, we evaluate the security of ECC masking with single and double shares that are considered secure when the random share r is chosen in a way that $\|r\| \geq \|n\|/2$ (see [RLMI21, RIL20, GRV17]) where $\|n\|$ stands for the bit-length of a natural number n . These implementations include hardware-accelerated constant time scalar multiplication (CM0), additive masking (CM1), multiplicative masking (CM2), and a combination of the previous two (CM3). For details, see Section 6.1.

2.3 Deep Learning

Throughout the paper, we assume a certain familiarity with standard deep-learning terms such as layer, activation function, and loss. Those terms are defined in widely available textbooks (e.g., [GBC16] and [Cho21]).

To process long traces efficiently GPAM borrows ideas from the recent advances in image patchification techniques which were introduced in vision transformers to efficiently process images [LMW⁺22].

In terms of architecture, GPAM leverages the *Transformer* architecture [VSP⁺17] which is at the heart of the recent breakthroughs in deep-learning including large language models (LLMs) such as chatGPT and Gemini. What makes the transformer architecture well-suited to side-channels attacks is its use of *self-attention* [VSP⁺17], which enables the model to efficiently understand long-range dependencies and capture contextual information. These abilities are key to build a generic and efficient side-channel attack model as it allows it to exploit complex data leaks that occur through interconnected relationships between distant data points. In our work, we use a high-performance transformer block called GAU (Gated Attention Unit) which was introduced by Hua et al. [HDLL22]. GAUs enhance the Transformer encoder block by replacing the vanilla attention and feed-forward network with a combined gating mechanisms that improve data representation and computation speed, leading to faster training and improved accuracy.

Activation functions play a key role in model accuracy by introducing different form of non-linearity. Different functions are better suited to different type of data and use-cases. As suggested by [XTG⁺20], we use swish [RZL17], a smooth activation function which is more robust in the presence of counter-measures. Last but not least GPAM heavily relies on using multi-task learning [Car98] to converge and generalize.

3 Threat model

We assume the attacker has access to a clone of the targeted hardware and the tools to collect power traces following side channel attacks' standard assumptions [CRR03].

Our main threat model is the white-box model, even though, as illustrated in Section 6.8, it is also able to perform well in a black-box setting. Following SCA standard model, we also assume attacks are carried in two phases: the *training* phase during which the attacker collects data using the cloned hardware to train their attacks and the *attack* phase where the attacker is attempting to recover secret from the targeted device.

During the attack phase, the model exclusively processes the raw traces from the targeted device to recover the targeted secret without any access to the countermeasure parameters regardless of the threat model considered. We ensure our dataset collection process is consistent with this modus operandi by using two different chips to collect our data: one chip is used for creating the training and testing data while the other one is used to collect the holdout dataset used for attack evaluations.

During training, we consider two threat models:

1. **Black-box threat model:** In this threat model, the attacker has no knowledge of the deployed countermeasures, only knows the input and output of the cryptographic primitive and can only control the inputs (chosen text attack).
2. **White-box threat model:** In this threat model, the attacker has full knowledge of the countermeasures used and they control all the protection parameters during the training phase. This is the main threat model used in this paper because it mimics the level of access that chip development teams or certification evaluators have.

4 Related Work

Machine-learning (ML) side-channel attacks. Machine learning has been repeatedly shown to be an effective approach to SCA. For example, Lerman et al. [LBM15] outperformed template attacks [CRR03] in recovering keys from masked implementations of AES, leveraging classical ML algorithms such as support–vector machine (SVM). Maghrebi et al. [MPP16] later applied deep–learning algorithms, including CNNs and LSTMs, to attack AES. Bursztein et al. [B⁺19, BP19] then proved the feasibility of full-trace attacks using deep learning. These attacks tend to require a higher number of traces than classical attacks, though other researchers also adopted MLPs and CNNs to reduce the number of traces required [CCC⁺19, ZBHV21, CDP17].

To improve upon these early results, research is still needed to overcome the following three challenges:

(1) **Trace preprocessing:** This remains a costly endeavor done by experts. There are works that develop techniques to address this issue. Notably, Won et al. [WHJ⁺21] developed a framework based on a multi-scale CNN to enable the integration of user-defined preprocessing phases. Hettwer et al. [HGG19] explored various image–classification metrics for finding points–of–interest in the signal. Zhou and Standaert [ZS20] proposed a technique based on residual networks for aligning SCA traces. Wu and Picek [WP20] used autoencoders to filter out noise added by mitigations such as clock jitter and random delays. Transformations of one dimensional traces into two dimensional images and using established network architectures for images have been studied by [HHGG20].

Direct use of the whole traces remains rare when the traces are long. Even very recent publications [HCM24] suggest that using raw traces of tens of thousands of points is still not a solved problem. To the best of our knowledge, the following are the only papers which directly target traces of at least 100k points using ML (the threshold is somewhat arbitrary since for each threshold there are papers which almost make it, e.g., [GJS20]

with 65k point traces). [MBC⁺20] target AES implementations automatically protected by code polymorphism (traces up to 160k points). Lu et al. [LZC⁺21] developed an ML architecture (autoencoders and attention mechanism) acting directly on raw traces of up to 300k samples to target AES implementations from public datasets. [GLS22] directly use traces of length up to 219k samples from the CHES 2020 contest. Our model has improved the result of [LZC⁺21] on the ASCADv1 variable key dataset. Our approach improves on prior art as it does not require trace preprocessing and can support very long traces, up to 16 million samples and up to 1 million points on a public dataset.

(2) Generalizability: Prior art has mainly focused on identifying the optimal network architecture for each device, implementation, and crypto algorithm [PP21, WAGP20]. This is an effective strategy to find an optimal solution for a specific attack configuration, but it is not clear how well it serves embedded engineers trying to identify leaks in a new product. Some works are addressing this issue, by searching for the right ML architecture based on various tools such as Information Theory [AGF23], Bayesian Optimization [WPP22], and Gradient Visualization [ZBHV20]. Pernin et al. [PCP20] take a different approach, using an ensemble of ML models based on average class probabilities to improve generalization. Whereas there are works that target multiple implementations (e.g., [WCPB21]), to the best of our knowledge, no prior work has studied generalization across multiple algorithms (e.g., AES, ECC, RSA). Our approach differs from prior art as we find a single architecture capable of generalizing across devices (with identical model weights), implementations, and algorithms (using the same tunable architecture), thus reducing training costs and heading toward a fully-automated SCA leakage evaluation for hardware certifications.

(3) Portability: Identical hardware devices, even when originating from the same production line, exhibit minute physical variations that result in differences in their power traces. ML models that generalize across devices are superior, as during training, they do not need access to the devices they will eventually attack. Prior research has looked into incorporating device-to-device variation into SCAAML training [RBA20, ZSX⁺20, BCH⁺20]. Our holdout datasets are also captured on different physical chips.

Side-channel attacks on ECC. Single-trace side-channel attacks on ECC aim to recover most of the secret bits in one execution of scalar multiplication. This is ideal for signature schemes, like ECDSA, which performs scalar multiplications on a fresh integer every time. However, these attacks [SI11, JB17, NC18, HIM⁺14, WPB19] have only been successful when scalar blinding has low entropy. Most recently, Pernin et al. [PCBP21] showed ML’s effectiveness in unsupervised attacks, recovering 90% of the secret bits when scalar blinding is performed with 32 and 64 randomly-generated bits. [RIL20] leveraged ML to attack ECC key generation by collecting multiple traces from scalar multiplication for the same secret.

Prior work has also shown lattice-based attack’s effectiveness in recovering keys from partial leakage collected on real cryptographic chips when no masking countermeasure is applied [MSEH20, JSSS20, RLMI21]. However, the effectiveness of SCAAML in assisting lattice attacks and bypassing stronger countermeasures is unknown. Goudarzi et al. [GRV17] combine lattice attacks with side-channel (using hypothetical SNR analysis) and show that attacks are probable when $\|r\|$ is 16, 32, 64 bits. Based on these works, the current understanding in the community is that if $\|r\| \geq \|n\|/2$, the countermeasure is safe.

To the best of our knowledge, our approach is the first to show that SCAAML driven lattice-attacks can recover the ECDSA key when $\|r\| \geq \|n\|/2$ ($\|r\| \geq 128$ bits in our case) and even for higher-order masking with more than one share.

5 GPAM

In this section, we present the GPAM (Generic Power Analysis Model) architecture and discuss how we train it. We start by discussing the training objective along with the metrics used to evaluate convergence. Next, we detail GPAM model architecture. Finally, we discuss relevant implementation details.

5.1 Training objective and metrics

Training objective GPAM’s training objective is to predict the value of a specific key byte k_i . Following standard machine learning practices, we cast this problem as a classification problem where the model is tasked to produce the correct byte value out of 256 possibilities. In practice, this translates to the model outputting softmax probabilities P for every possible key candidate c , i.e., $P[k_i = c]$ for $c = 0x00, \dots, 0xFF$, each indicating the likelihood that the predicted key byte k_i is equal to c . We use categorical cross-entropy as our loss function.

Metrics We use the following metrics to evaluate GPAM performance:

- **Accuracy** is our main metric. It is defined as the categorical accuracy of the output, meaning the ratio of which the model predicts the correct output. The baseline accuracy for a random guess is $1/256 = 0.39\%$. We will indicate it in the rest of the paper with the 🎯 symbol.
- **Rank** is the position of the correct value in the ranking of predicted byte values, sorted in descending order by probability. A rank of zero is assigned to the highest probability and 255 is assigned to the lowest probability, given that a byte has 256 possible values.
- **MaxRank** is the maximum rank over a set of model predictions for a batch of examples. The baseline MaxRank is 255. A value less than 255 implies that the key space required to bruteforce the correct value, in the worst case, has been successfully reduced, as each correct prediction is contained in a smaller range of values.
- **MeanRank** is the average rank of predictions. The baseline MeanRank for a random guess is 127.5. A lower value implies that the key space was successfully reduced and that on average, the correct value is within a smaller range of values.
- **Confidence** reflects the standard margin sampling confidence [BJSR22] the difference between the highest and second-highest probabilities (i.e., the value `prediction.sort(); prediction[-1] - prediction[-2]`). Intuitively, this metric is one of the most informative way to measure how much the model is confident that the predicted value is correct.

These metrics are well suited to inform an evaluator about the presence of leakage. For instance, one such indicator is the accuracy rising over the threshold given by the 3σ rule. However, the leakage detected by these metrics does not imply a successful attack. To account for this, we also evaluate models as part of end-to-end attacks from trace collection to recovered key to understand their real-world performance. For example, we attack ECDSA end-to-end in Section 6.8, and AES in Section 7.1.

Following machine learning best practices, we conduct experiments with the GPAM model on the test split (called validation split by some authors), then pick the best model according to the metrics discussed above, and evaluate only once on the holdout (which is collected on a different chip) by carrying the end to end attack.

5.2 Architecture

At a high-level, GPAM is composed of three functional components, as depicted in Figure 1: a *temporal patchification* stem designed to group traces into a sequence that preserves the

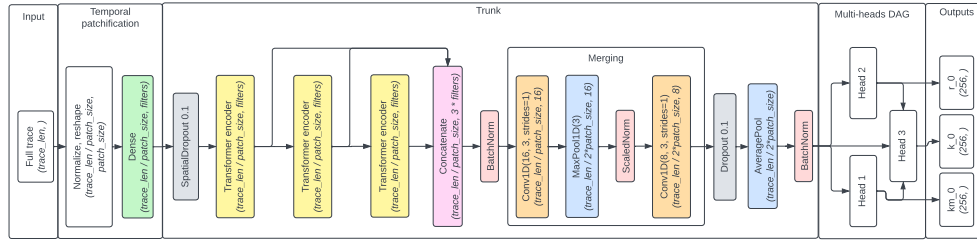


Figure 1: GPAM architecture for predicting k_0 in CM1 (see Section 6.1 for description of CM1 and the attack points k_0, km_0, r_0). The attacked key byte is k_0 , with k_0 having related outputs km_0 and r_0 .

temporal inductive bias, a *trunk* that attends to the temporal sequence using transformer encoder blocks to extract information, and a multi-headed directed acyclic graph that is used to implement multi-task learning, i.e., predicting multiple values at once.

5.2.1 Temporal Patchification

The temporal patchification stem has two goals:

1. Preserve the temporal inductive bias while making the sequence faster to process by transformer encoder blocks by grouping adjacent points – recall here that attention computation cost is quadratic in the length of its input sequence. To achieve this we reshape the trace into blocks of N contiguous non-overlapping chunks, or “patches”. This approach, while slightly different, is inspired by state-of-the-art image patchification techniques [LMW⁺22].
2. Potentially provide positional information to allow the transformer encoder block to perform efficiently. This is done by injecting global positional encoding information to the sequence [CTB⁺21].

5.2.2 Trunk

The trunk’s main function is to attend to the patchified sequence by extracting the latent representation of the traces needed by the heads to predict the targeted values. To do so, GPAM uses a trunk made of three state-of-the-art GAU transformer encoder blocks [HDLL22] that are able to isolate and process long-range interacting features. In addition to the transformer blocks, the trunk also includes a combiner module made of convolutional layers that is meant to combine the output of the three encoder blocks into a unified latent representation. Combining the output of the encoder blocks instead of using the output of the last one, as traditionally done in NLP, is useful because each block extracts features at a different level of “abstraction”. Those multi-level representations are commonly used in other signal processing applications such as speech recognition [CZH⁺21].

5.2.3 Heads and Relational Outputs

The last component of GPAM is its multi-headed DAG (directed acyclic graph). This component is designed to achieve two goals:

1. **Allow multi-task learning:** GPAM relies on multi-task learning [Car98] to perform efficiently against masked implementation, as reported in Section 6.4. Multi-task learning is accomplished by not only predicting the targeted byte value but also predicting intermediate values such as the mask and random nonce values.
2. **Inject domain expertise:** Standard multi-task learning involves jointly predicting

values without establishing relation between the outputs. We found out, as reported in Section 6.4, that we can increase GPAM performance by representing the outputs as a DAG where intermediate outputs feed into the byte prediction output, as depicted in Figure 1. This allows the model to benefit from expert understanding and makes it easier for it to learn which intermediate values are useful for computing a given output. We note that defining those relations is fully configuration driven and does not require to change the model architecture or fiddle with the code.

5.2.4 Heads design

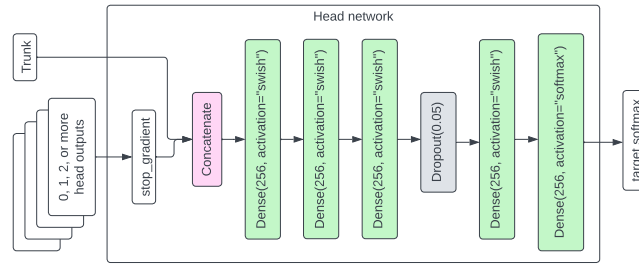


Figure 2: Single head output.

Unlike standard transformer architectures, where the output is a single layer, we discovered during our architecture search that having a deeper head architecture improves model performance. As visible in Figure 2, GPAM head architecture comprises several dense layers and a single dropout layer. Extensive initial testing during the architectural development phase revealed that adding normalization layers, residual connections, or more dropout layers did not seem to improve model performance or convergence speed.

5.3 Hyper-parameter tuning

GPAM is designed to be automatically hyper-tuned (using `Keras-tuner`) to quickly adapt to new cryptographic implementations. To minimize GPAM production costs, we focus on reducing the tuner search space to a minimum. We perform a one-time extensive architecture search to isolate which parameters should be hyper-tuned and which should be considered canonical. We emphasize that this search is a one-time cost paid by this research, and will not be run when deployed in an automated testing environment. For example, the activation function used (Swish), the number of layers per head, and the number of GAU blocks (3) all proved to be close to optimal choices across algorithms and implementations and therefore were excluded from hyper-tuning.

Table 1: Model hyper-parameters for each targeted implementation used in this paper.

	CM0	CM1	CM2	CM3	ASCADv2
Batch size	128	64	64	32	64
Steps per epoch	200	200	200	400	1,000
Epochs	25	500	500	500	150
Target learning rate	0.0006	0.0006	0.0006	0.0003	0.00005
Merge filter 1	16	16	16	16	0
Merge filter 2	8	8	8	8	0
Trace length	1,620,000	4,194,304	8,388,608	16,777,216	1,000,000
Patch size	1,200	2,048	4,096	4,096	400

All in all, at the end of the architecture search, GPAM only requires 8 parameters to be tuned for each new application. The values of those 8 parameters for all the ECC and AES implementations targeted in this paper are reported in Table 1.

- **Batch size** depends purely on the hardware available for training, and should be

maximized as much as the GPU memory allows (as a standard machine learning practice, doubling `batch size` implies halving `steps per epoch` and doubling `target learning rate`).

- `Trace length` and `patch size` depend on the size of the dataset, and the capture sample rate. The Patch size is the square root of trace length (rounded).
- `Target learning rate` and `number of epochs` depends on the difficulty for the optimizer to learn. Learning rate is to be searched for in logarithmic scale (good values are 0.001, 0.0001, 0.00001).
- `Merge filter 1` and `Merge filter 2` are used to make the trunk output length more manageable. Merge filter 1 and 2 are searched among the powers of two, and as a rule of thumb merge filter 2 is half of the merge filter 1.

We emphase that hyper-tuning the 8 parameters is straightforward and automated. For example, we hyper-tuned GPAM to attack ASCADv2 using NVIDIA RTX 4090 GPUs. The training totaled 29 days (cumulative across GPUs), which amounts to 7 hours spent training for each parameter configuration. We note here that during hyper-tuning we do not train fully, instead we only use 10% to 30% of the number of epochs used for the final full training. This amount of compute being enough to figure out which parameter values are the best.

Using a server with eight GPUs, a standard number when using NVIDIA SMX technology, this hyper-tuning search can be done under 3.6 days. For example running this training on a similar configuration on Google Cloud Compute costs under \$2.5k³. This hypertuning can alternatively be performed on a local server with equivalent features if a local deployment is preferred, albeit for a higher startup cost. This is a comparatively small price to pay to avoid having experts manually preprocess traces and manually tweak attack parameters, whose cumulative hourly rate can quickly surpass this dollar value. Also, note that this hypertuning price has to be paid only once per targeted combination of hardware platform, algorithm, and countermeasure.

5.3.1 Weight Initialization Influence

Training instability is a well-known issue in deep learning in general, and for transformers in particular, where an unlucky initialization can lead to model collapse or sub-par performance (see for instance [NS19]). Following machine-learning practices, we mitigate this issue by using a custom learning-rate scheduler. We start from a low initial learning rate value, then steadily increase it to reach its target value, and finally decrease it with a cosine decay as the model converges [LH16]. During the warm-up phase, a low learning rate allows the model to perform smaller steps along the gradient, reducing the influence of the weight initialization.

5.4 Implementation

We implement the GPAM architecture and conduct training using TensorFlow [AAB⁺15] and the Keras API [C⁺15]. We use Keras Tuner [OBL⁺19] to hypertune GPAM automatically. The temporal patchification code was specially designed and implemented for GPAM. The GAU layer is a custom implementation based on the pseudo-code provided in [HDLL22]. The relational output technique is our own contribution and its implementation unique to GPAM. All in all, the GPAM code is about 1,000 lines of Python.

The estimated training times for attacking the various datasets used in the paper, using

³This price is calculated in September 2023 with Google Cloud Pricing Calculator (<https://cloud.google.com/products/calculator>) running a `a2-highgpu-8g` host for 3.6 days. In reality, the GPUs offered by that configuration (A100 80GB) are slightly more performant than our 4090 cards (benchmark <https://lambdalabs.com/gpu-benchmarks>), so this price tag estimate is an upper bound.

Table 2: Training time estimates.

Dataset	CM0	CM1	CM2	CM3	ASCADv2
Training time [hours]	1	24	48	150	20

an NVIDIA RTX 4090 as reference GPU, are reported in Table 2. In practice, we use multiple servers with various GPU configurations, as the total computation required to complete all our experiments presented in this paper requires over 1 year of computation. As mentioned throughout the paper, we did our best to avoid running unnecessary experiments to minimize carbon emissions.

6 Generic attacks against hardware-protected ECC

In this section, we evaluate GPAM’s ability to generalize to multiple hardware-protected implementations by performing side-channel attacks against scalar multiplication on four distinct implementations of ECDSA. The targeted hardware implementation includes a constant-time implementation and three distinct algebraic masking implementations that are considered state-of-the-art protections [Cor99, BR23]. We start by describing the implementations targeted, then describe our collection process. Next, we evaluate GPAM’s performance against those datasets. Finally, we discuss how we can attack the ECDSA signature scheme by combining the partial nonce K recovered via GPAM and a lattice attack.

Here we answer the following questions:

1. Does GPAM generalize to multiple hardware implementations? (Section 6.3)
2. Is Multi-task learning needed to detect leakage in a protected implementation, and if yes, which tasks are needed? (Section 6.4 and Section 6.5)
3. Are related outputs only learning a function of other outputs, or are they also using the latent representation provided by the trunk? (Section 6.6)
4. How many traces are needed for GPAM to successfully detect leakage in various implementations? (Section 6.2)

6.1 Targeted hardware implementations

Given our goal to evaluate GPAM against highly-protected hardware implementations, we use the NXP K82F dedicated cryptographic accelerator (LTC – LP Trusted Crypto) as a base for all implementations to perform constant-time hardware-accelerated scalar multiplication and point addition. Relying on this accelerator ensures that all our implementations are not vulnerable to timing attacks or software-based leakages. All our ECDSA implementations use the elliptic curve FRP256⁴ from ANSSI, but our results apply equally to other curves (e.g., NIST P-256), as the scalar multiplication algorithm is typically the same for all Weierstrass curve implementations.

6.1.1 Countermeasures

We evaluate the following four implementations to highlight the effectiveness of GPAM against increasingly stronger protections. Here, we use \square to denote computation done in software and \blacksquare to indicate computation done on the chip. The $k \leftarrow \delta_L(N)$ notation indicates the generation of a random number with N bits where each bit is selected independently and uniformly at random. The four implementations considered in this section are:

1. **Constant-time execution (CM0).** A simple countermeasure effective against

⁴<https://neuromancer.sk/std/anssi/FRP256v1>

timing attacks, but not power side channels. It exclusively relies on our chip’s constant-time accelerated scalar multiplication, without randomizing the secret multiplier.

2. **Additive masking (CM1):** This implementation is significantly more resistant to power side channel attacks compared to CM0, thanks to the addition of multiplier masking. A random integer r is added to k so the scalar multiplication executes on the blinded secret scalar. More formally, it

- (a) Chooses an independent 256-bit random mask r .
- (b) Computes the difference km between the secret multiplier k (the ECDSA nonce) and the mask r .
- (c) On chip, computes $P_{km} = km \times G$ and $P_r = r \times G$.
- (d) On chip, computes $P_{km} + P_r$ and returns this value.

Here is the pseudo code used to implement this scheme:

$k \leftarrow \delta_L(256)$	(secret multiplier ☐)
$r \leftarrow \delta_L(256)$	(random mask ☐)
$km = (k - r) \bmod n$	(k masked ☐)
$P_{km} = km \times G$	(scalar multiplication ⚡)
$P_r = r \times G$	(scalar multiplication ⚡)
$Result = P_{km} + P_r$	(equal to $k \times G$ ⚡)

3. **Multiplicative masking (CM2):** In this implementation, we use a Euclidean division rather than addition computation, which is another canonical way to mask the secret scalar [BR23, Cor99]. Formally, it:

- (a) Chooses an independent 128-bit random mask r .
- (b) Computes km , the quotient of the division of the secret multiplier k and r . It also computes the *remainder*.
- (c) On chip, computes $P_{km} = km \times G$ and $P_r = rem \times G$.
- (d) On chip, computes $P_{km2} = r \times P_{km}$, then returns $P_r + P_{km2}$ (equal to $k \times G$).

$k \leftarrow \delta_L(256)$	(secret multiplier ☐)
$r \leftarrow \delta_L(128)$	(random mask ☐)
$km = \lfloor k/r \rfloor$	(☐)
$rem = k \bmod r$	(☐)
$P_{km} = km \times G$	(⚡)
$P_r = rem \times G$	(⚡)
$P_{km2} = r \times P_{km}$	(⚡)
$Result = P_r + P_{km2}$	(equal to $k \times G$ ⚡)

4. **Combined countermeasure (CM3):** This combines CM1 and CM2 techniques in an attempt to further increase security with higher-order masking.

$k \leftarrow \delta_L(256)$	(secret multiplier ☐)	$P_{km1} = P_{r2} + P_{km2}$	(⚡)
$r_1 \leftarrow \delta_L(256)$	(CM1 random mask ☐)	(CM2 (b) instead of $P_{r1} = r_1 \times G$)	
$r_2 \leftarrow \delta_L(128)$	(CM2 random mask ☐)	$km_3 = \lfloor r_1/r_3 \rfloor$	(☐)
$r_3 \leftarrow \delta_L(128)$	(CM2 random mask ☐)	$rem_3 = r_1 \bmod r_3$	(☐)
$km_1 = (k - r_1) \bmod n$	(☐)	$P_{km3} = km_3 \times G$	(⚡)
(CM2 (a) instead of $P_{km1} = km_1 \times G$)		$P_{km3} = r_3 \times P_{km3}$	(⚡)
$km_2 = \lfloor km_1/r_2 \rfloor$	(☐)	$P_{r3} = rem_3 \times G$	(⚡)
$rem_2 = km_1 \bmod r_2$	(☐)	$P_{r1} = P_{r3} + P_{km3}$	(⚡)
$P_{km2} = km_2 \times G$	(⚡)	$Result = P_{km1} + P_{r1}$	(same as $k \times G$ ⚡)
$P_{km2} = r_2 \times P_{km2}$	(⚡)		
$P_{r2} = rem_2 \times G$	(⚡)		

Random masks selection We discuss the choice of the random mask r (or r_1, r_2, r_3 in the case of CM3). For additive masks, we choose r in a way that $\|r\| = \|n\|$, i.e., 256 bits. However for the multiplicative masks, we have to choose r in a way that $\|r\| < \|k\|$, otherwise, the mask would not be effective. We choose r as $\|r\| = \|n\|/2$, i.e., 128 bits, to ensure that it meets this property, but it also achieves the level of resistance to side-channel attacks, as suggested by prior work [GRV17, RLMI21]. We also choose each byte of k and r (or r_1, r_2, r_3 for CM3) independently and uniformly at random for each computation to ensure we are performing attacks against implementations that do not have a bias in their random entropy.

6.2 Power trace collection

Our capture setup consists of a Chipwhisperer CW308 UFO board with a CW308T-K82F target board connected to it. The firmware of the target chip was solely responsible for curve addition and scalar multiplication. We did not rely on any software implementation for curve arithmetic. Scalar multiplication and point addition operations were performed by sending an integer and a point, or two points, respectively, to the LTC. For creating labels for model training, we additionally record on the host computer the secret multiplier and random parameters used for masking each trace.

Capture setup We collect power measurements using the *Teledyne LeCroy WavePro 404HD-MS* [Tel] oscilloscope connected to the embedded resistor shunt on the *ChipWhisper NAE-CW308T-K82F* [New] target board. The oscilloscope probe is hooked to the test point TP5 of the CW308 UFO board to measure the current, while digital channel D0 is connected to the GPIO4/TRIGGER pin to get the trigger signal, which starts the capture. We insert a 7.37MHz crystal into the X1 socket and adjust the clock source selection jumper J3 to the CRYSTAL position to provide the clock signal. This configuration ensures that there is no correlation between the target chip clock and the oscilloscope sampling clock, resulting in asynchronous measurements. The oscilloscope channel is set to AC coupling with a bandwidth limited to 200MHz. The first scalar multiplication is always aligned using a trigger signal for each operation, and there is no additional alignment performed. The trigger signal was configured to stay high during each operation performed by the LTC. We use the first rising edge of the trigger signal as the oscilloscope trigger to start capturing.

Experimental leakages We ensure that no UART communication is leaking by conducting one experiment where we replace captured points in the training set with Gaussian noise when the trigger signal is low. After training a model, we observe no performance loss when compared to training on raw traces, indicating that no discernible UART leakage is occurring in the replaced points. Note that all other experiments are conducted with raw traces.

Datasets Collected Table 3 provides a technical summary of the datasets generated using the implementation discussed in Section 6.1. We use the SCAAML ([B⁺19]) dataset library to store our traces and attack point values as TFRecord files. We ensure that no key is reused between the splits by tracking which keys were previously used. Overall each dataset collection process takes anywhere from several weeks (CM0) to months (CM3) to complete.

Note that to ensure that the attack generalizes between chips of the same family despite potential subtle hardware variations, we use a different chip to collect the training/test splits and the holdout split. The holdout splits, following machine learning best practices, are never used to tune the models or during experimentation. Instead, they were reserved for the final evaluation presented in Section 6.3.

Note that Table 3 also reports the ASCADv2 dataset that we use to evaluate GPAM

generalization across multiple algorithms in Section 7. This dataset was made public in [ERR⁺18]; we simply convert it to the SCAAML dataset format. Since there was no apparent restriction on how to divide the samples into splits (train, test, holdout), we took a portion of consecutive samples for train, a portion for test, and a portion for holdout.

Table 3: List of ECC evaluation datasets used in this study to evaluate GPAM generalization to multiple hardware implementations. The names refer to protected implementations described in Section 6.1. The table also includes the ASCADv2 dataset collected in [ERR⁺18] that is used in Section 7 to evaluate GPAM generality across multiple algorithms.

Name	Trace length	Train length	Test length	Holdout length	File size [TB]
ECC CM0	1,6M	57,344	8,192	8,192	0.2
ECC CM1	5M	194,544	8,192	8,192	1.5
ECC CM2	10M	122,880	8,192	8,192	2.1
ECC CM3	17,5M	122,880	8,192	8,192	3.7
ASCADv2	1M	640,000	80,000	80,000	0.9

6.3 Generalization over multiple implementations

Overall, GPAM can successfully attack all four ECC hardware implementations in white-box settings using multi-task relational outputs training as reported in Table 4. These results are computed on the holdout splits which, as discussed previously, are captured on a different chip of the same family and were not used for model tuning or any other experiments discussed later in this section. As discussed in the threat model section (Section 3), white-box setting means that the model had access to the intermediate values (masks and random values) during training.


Note that we only evaluate GPAM on the initial (most-significant byte k_0), middle (k_{15}), and last (least-significant byte k_{31}) byte of each implementation, as those bytes are representative of GPAM performance against those implementations. We make this choice because these experiments are resource intensive, and the goal of GPAM is to identify leakage.

Table 4: GPAM white-box key byte recovery success rate on the four ECC hardware protected implementations holdout splits.

Dataset	Attack point	Accuracy [%]	MeanRank	MaxRank
CM0	k_0	100.00	0	0
CM0	k_{15}	100.00	0	0
CM0	k_{31}	100.00	0	0
CM1	k_0	78.80	0.75	192
CM1	k_{15}	93.20	0.31	253
CM1	k_{31}	92.98	0.24	218
CM2	k_0	66.22	1.40	254
CM2	k_{15}	0.30 \boxtimes	127.29	255
CM2	k_{31}	11.31	8.76	233
CM3	k_0	8.60	19.77	255
CM3	k_{15}	-	-	-
CM3	k_{31}	0.37 \boxtimes	127.89	255

As expected, as the strength of the protection increases, the model accuracy decreases to the point where for CM3, only the initial byte can be attacked successfully. Note that an accuracy of 0.4% is close to random chance. We hypothesize that increasing GPAM’s performance against stronger countermeasures requires more training data, not increased model capacity. This is empirically supported by our experiment in Section 6.7, which looks at model accuracy as a function of the number of training traces, showing that

Table 5: GPAM black-box key byte recovery success rate on the first three ECC hardware protected implementations using the holdout splits.

Dataset	Attack point	Accuracy [%]	MeanRank	MaxRank
CM0	k_0	100	0	0
CM1	k_0	0.29 	127.5	255
CM2	k_0	22.81	5.55	231

GPAM’s accuracy against CM3 only starts to rise past 100k traces. Furthermore, looking at the MaxRank results, it is clear that the model did not fully generalize for CM1, CM2 and CM3, as it is close to its 255 upper bound.

In Table 5, we report GPAM results under black-box settings on the holdout splits. Unlike when using the white-box settings, GPAM is not always able to successfully recover even the initial byte k_0 . Interestingly, GPAM fails to recover CM1 k_0 but is able to recover CM2 k_0 , which is surprising given that in the white-box setting, CM1 appears to be easier than CM2. We hypothesize that CM1 is harder in the black-box setting because its random mask uses 256 bits whereas CM2 only uses 128 bits. Having access to intermediate values seems to make the size of the random mask irrelevant in the white box setting but a strong defense in the black box setting.

6.4 Multi-task effectiveness evaluation

In the following set of experiments, we study whether using multi-task learning improves GPAM’s accuracy. In particular, we are interested in understanding which additional tasks beyond the key byte prediction improve model accuracy, if any. To not taint our holdout splits, the results reported in this section are computed on the test splits. Once again we only perform experiments on representative bytes to limit the computation time, namely the initial bytes (k_0, k_1, k_2), middle one (k_{15}) and final ones (k_{29}, k_{30}, k_{31}). For that same reason, we also only target the middle of the road dataset CM2 and reserve the study of CM1 for the ablation study discussed later in the paper in Section 6.5.

Overall, there are two types of additional tasks that can be included in the training: *Adjacency predictions* and *Intermediate predictions*. *Adjacency predictions* ask the model to predict the key bytes on the left and the right of the targeted bytes with the hope it helps with carry issues and generalization. *Intermediate predictions* involve the model predicting the value of intermediate computation points, including mask and random nonces values.

The model can operate in two modes when performing multi-task learning using intermediate values: the *multi-outputs* mode and *relational outputs* mode. In the *multi-outputs* mode, the model outputs all the asked values without any interaction between outputs. This is the classical form of multi-task learning used by many models to boost generality and accuracy. In the *relational output* mode, as illustrated in Figure 1, we create a directed acyclic graph between the heads to model expert knowledge of how the outputs relate to each other according to the protecting algorithm (CM1-CM3). Obviously this type of knowledge is only available in white box testing conditions.

Notation To make the results tables easier to understand, we are using the following visual convention to distinguish between the various relation conditions:

- Circles are byte indexes centered at the column index (if the column byte index is i then the circles represent $[i - 1, i, i + 1]$).
- A white circle \circ at position j means not leveraging multi-task learning.
- A gray circle \ominus at position j means using multi-task learning.
- A black circle \bullet means using relational outputs learning.

Here are a few examples of such notations for the column k_2 of Table 6:

- ○ ○ Outputs: k_2 , relations: []. Byte k_2 is targeted and no multi-task learning is used.
- ● ○ Outputs: k_2, km_2, r_2 , relations []. The model operates in the multi-task learning mode and predicts the values of k_2, km_2 , and r_2 .
- ● ○ Outputs: k_2, km_2, r_2 , relations [$km_2 \rightarrow k_2, r_2 \rightarrow k_2$]. The model uses relational outputs to predict the values of k_2, km_2 , and r_2 . The values of km_2 and r_2 are fed into k_2 .
- ● ○ Outputs: $k_2, km_1, r_1, km_2, r_2$, relations [$km_1 \rightarrow k_2, r_1 \rightarrow k_2, km_2 \rightarrow k_2, r_2 \rightarrow k_2$]. The model uses relational outputs and adjacency relations to predict the values of k_2, km_2 , and r_2 . The values of km_1, r_1, km_2 and r_2 are fed into k_2 .

Table 6: GPAM accuracy in % on CM1 test dataset when trained using various forms of multi-task learning.

Relations	k_0	k_1	k_2	k_{15}	k_{29}	k_{30}	k_{31}
○ ○ ○	0.39	0.39	0.20	0.20	0.39	0.78	0.59
○ ● ○	81.50	79.20	85.45	43.36	86.62	86.13	92.68
○ ● ○	85.35	71.88	88.09	85.64	85.64	87.21	92.58
● ● ○	–	67.68	81.45	40.82	82.52	83.38	93.65
○ ● ●	87.40	74.80	83.98	63.67	82.03	87.70	–
● ● ●	–	63.38	80.96	41.11	84.08	88.48	–

Results Overall, we observe that multi-task learning is needed for the attack to succeed as reported on CM1 in Table 6. Without any relations, denoted using the ○ symbol, the model predictions are unable to exceed random chance ($k_0, k_1, k_2, k_{15}, k_{29}$) or barely exceed it (k_{30} and k_{31}). Using the simplest form of multi-task learning, denoted using the ● symbol, the model obtains high accuracy for almost all the bytes except k_{15} . Using relational-output, denoted using the ● symbol, the model accuracy improves further overall compared to using multi-task learning.

Conversely, the effectiveness of adjacency relations is marginal. As reported in Table 6, adjacency relations introduce model instability with the accuracy slightly increasing on some bytes (e.g., k_0, k_{30}, k_{31}) but decreasing significantly on others (e.g., k_{15}). Given our goal to have a stable fully automated attack, we decided against using adjacency relations. We leave it as future work to make better use of them.

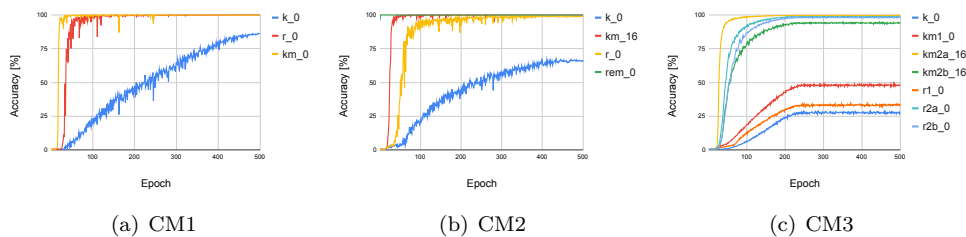


Figure 3: ECC validation accuracy for k_0 when using relational outputs.

Effect of multi-task learning on model convergence The positive impact of multi-task learning is best visualized by looking at how each of the output accuracies improve as training progresses. Regardless of the hardware implementation, we observe that outputs start to converge one after the other. For example, as visible in Figure 3(a), we observe that the mask prediction (km_0) accuracy rises before the random nonce (r_0) prediction accuracy improves, and the key value (k_0) prediction starts converging only after both the mask and the random nonce have reached a high accuracy. The same effect is observed for CM2 (Figure 3(b)) and CM3 (Figure 3(c)).

Additionally, we observe that in each case the model first learns to predict the mask

values and then the random nonces. This behavior is consistent with the hypothesis that multi-task learning is critical to create generalized SCAAML attacks as it allows models to learn to "unpack" protections one step at a time. It also supports the hypothesis that black-box attacks are significantly harder, because models greatly benefit from the extra information. Last but not least, this behavior seems to confirm the effectiveness of higher-order masks against advanced side channel attacks such as SCAAML.

6.5 Multi-task ablation study

In this section, we perform an ablation study to better understand which intermediate values are needed for the attacks to succeed on CM1 and CM2. We exclude CM0, as there are no intermediate values. We also exclude CM3, as GPAM's relatively low accuracy on this dataset makes it hard to confidently separate the results, and CM3 experiments would take roughly 16 months of computation.

Table 7: CM1 relational outputs ablation.

Target	Dependency	Accuracy [%]	MeanRank	MaxRank
k_0	km_0, r_0	86.26	0.17	9
k_0	km_0	0.39	130.00	255
k_0	r_0	86.60	0.14	4
k_0	Nothing	0.29	127.90	255
k_{15}	km_{15}, r_{15}	44.00	1.07	58
k_{15}	km_{15}	0.48	125.60	255
k_{15}	r_{15}	0.29	126.20	255
k_{15}	Nothing	0.48	125.69	255
k_{31}	km_{31}, r_{31}	92.87	0.09	15
k_{31}	km_{31}	0.48	123.46	255
k_{31}	r_{31}	93.85	0.07	7
k_{31}	Nothing	0.20	126.94	255

For CM1, as reported in Table 7, removing the prediction of the mask (r_*) at training time results in the model being unable to successfully attack CM1. Removing the prediction of the random nonce (km_*) drastically reduces the accuracy of the model for the middle key byte but has no effect on the initial and last byte. We are not sure why this happens.

Table 8: CM2 relational outputs ablation.

Target	Dependency	Accuracy [%]	MeanRank	MaxRank
k_0	Nothing	18.26	8.44	235
k_0	Outputs: r_0, km_{16}, rem_{16}	52.44	1.78	235
k_0	r_0, km_{16}, rem_{16}	66.50	1.99	252
k_0	r_0, km_{16}	66.50	2.14	254
k_0	r_0, rem_{16}	39.16	4.47	244
k_0	km_{16}, rem_{16}	32.40	4.21	223
k_0	r_0	39.26	4.43	247
k_0	km_{16}	36.23	2.63	160
k_0	rem_{16}	11.72	15.37	203
k_{31}	Nothing	0.39	125.60	255
k_{31}	Outputs: $r_{15}, km_{31}, rem_{31}$	9.00	6.74	192
k_{31}	$r_{15}, km_{31}, rem_{31}$	10.40	10.60	238
k_{31}	r_{15}, km_{31}	10.74	9.90	142
k_{31}	r_{15}, rem_{31}	8.49	10.33	207
k_{31}	km_{31}, rem_{31}	8.59	8.58	195
k_{31}	r_{15}	9.27	8.99	155
k_{31}	km_{31}	1.36	41.38	215
k_{31}	rem_{31}	8.59	9.90	165

We only target k_0 and k_{31} in Table 8, as the model is only able to target the most and least significant bytes of CM2.

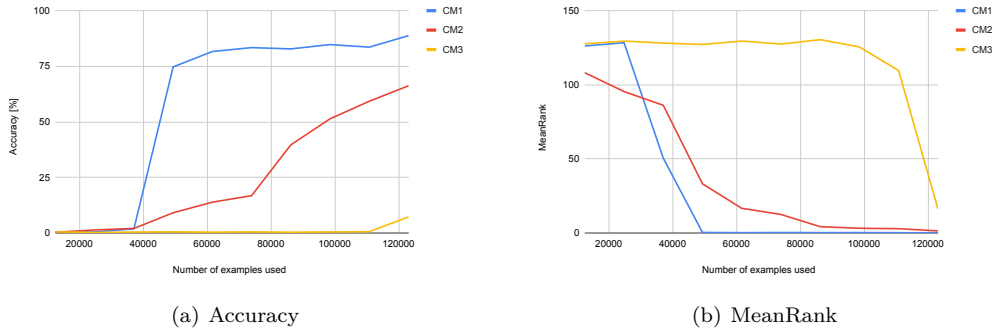


Figure 4: ECC k_0 accuracy and MeanRank while using only a part of the dataset. Results evaluated on the test split.

6.6 Trunk ablation study

In this ablation study, we evaluate whether the relational outputs benefit from the trunk output. To validate this hypothesis, we try to predict k_{15} but with only using the output of the km_{15} and r_{15} heads not the output of the trunk. Additionally we prevent the model from cheating and encoding k_{15} prediction information in the intermediate outputs by applying a stop gradient on the km_{15} and r_{15} output layers. For this ablation study, the head outputting k_{15} looks like the one in Figure 2 when one removes the “Trunk” input.

Table 9: CM1 ablation of trunk output for the k_{15} head.

Trunk	Accuracy [%]	MeanRank	MaxRank
yes	85.64	0.52	198
no	30.6	1.53	126

When the key byte prediction head is not directly connected to the trunk, the model exhibits an accuracy drop of 55 percentage points as reported in Table 9. Those results strongly support the hypothesis that the transformer encoder blocks’ latent representations and compute power are critical for accurate prediction.

6.7 Dataset size impact evaluation

In this section, following the insights of [KMH⁺20] where the authors showed that transformer models are bound by capacity, compute, or data, we attempt to determine which of these factors is limiting GPAM performance. We already know, thanks to the experiment ran in Section 6.4, that GPAM is not bounded by compute since the performance plateaus after a few hundred epochs - see Figure 3(a) for example. Accordingly, to decide whether the model is capped by the amount of data available or the model capacity/architecture, we trained the same model on an increasing number of traces from CM1, CM2, and CM3. We take 10%–100% of the 122,880 available examples and learn to predict k_0 .

We found out, as visible in Figure 4(a) and Figure 4(b), that GPAM is most likely bounded by the lack of data since the accuracy keeps rising as the amount of training examples increases. In particular, for CM3, GPAM starts to generalize only when at least 110,000 examples are used, suggesting that increasing the dataset size would most likely lead to significant accuracy gain. However, the CM3 dataset already requires 3.7TB of storage, so we decided against increasing the dataset size further, as GPAM is already able to successfully attack CM3, and increased accuracy does not bring significant additional

benefits.

6.8 ECDSA attack

Due to the presence of highly-secure countermeasures, GPAM is not able to recover all bytes of the multiplier at once with sufficient confidence. Thus, a single-trace attack similar to [WCPB21] seems out of reach. This is expected, as custom attack models have an edge over generalized models in this regard. However, a multi-trace attack is feasible, as we shall demonstrate. Specifically, the partial leakage we obtained can be combined with a lattice attack to recover the key from about 8000 traces, as demonstrated by [MSEH20, RLMI21].

We apply a standard lattice attack [HGS01, NS02] to leverage the partial leakage from ECC scalar multiplication protected by either CM1, CM2 (even blackbox), or CM3 and recover the private key from ECDSA ([JMV01]). To simulate a realistic attack, we take our holdout split (captured on a different physical chip than the training data), treat it as the nonce multiplication of ECDSA $k \times G$ explained in Section 2.1, and predict the most significant byte. One caveat is that our multipliers in the holdout split are chosen to test the deep learning model so that each byte is chosen independently and uniformly at random between 0 and 255. Since for ECDSA we require the nonce k to be $1 \leq k < n$, where n is the size of the elliptic curve, we exclude all measurements that have the multiplier outside of this range, after which we are left with roughly 7,800 examples (depending on the dataset – CM1, CM2, CM3). We try to closely simulate an attacker with a profiling device (i.e., the chip the attacker uses to collect training and validation splits) and traces from the device under attack. The attacker is free to leverage the model’s outputs as they wish, as long as they are able to recover the secret key in reasonable time. The way of using prediction confidence described below is specific to our attack configuration (model and dataset), and we do not claim that it directly transfers to other attack configurations.

The lattice attack requires all predictions to be correct for it to work, or as shown by [DDME⁺18], it can tolerate a very small number of noisy signatures. After trial and error with the predictions, our intuition is that if we turn byte predictions into predictions of the four most significant bits (MSBs), we will achieve the highest accuracy. To compute the probability of the 4 MSB of the nonce for a given signature, we sum the probabilities of the 16 possible LSB values with the same 4 MSBs. Table 10 shows the prediction accuracy of the four most significant bits. As we can see, we have a much higher accuracy in this case.

Table 10: Results of predicting the 4 most significant bits (on holdout).

Experiment	Accuracy [%]	MeanRank
CM1	94.83	0.06
CM2	96.39	0.07
CM2 black-box	85.75	0.20
CM3	71.86	0.87

A lattice attack using the four MSBs of the nonce requires 80 signatures [JSSS20, MSEH20] predicted by the model. Still, random sampling of 80 signatures will have a high chance of having erroneous signatures, especially when the accuracy is still under 90%. However, we notice that in cases where the correct key 4 MSBs are detected with high accuracy, the prediction value has a higher *confidence* (highest probability - second highest probability). We use the confidence of predictions as weights when randomly sampling (using the parameter `weights` of `random.choices` in Python). This approach retrieves the secret key after several retries for CM1 and CM2.

For the case of CM3, we employ the following heuristics to succeed in the attack. First, we give more weight to samples with higher confidence; we achieve this by using confidence to the power of eight as weights when sampling for the subset used in the lattice attack. The constant eight is chosen by profiling the attack on the validation set (also roughly

8,200 examples). Second, we discard samples with too high confidence (more than 0.25, chosen by trial and error) when predicting the byte value, discarding roughly 10% of examples. This excludes all (and thus also the wrong) predictions that are too confident and would occur in most random samples.

We use the lattice construction of [BVdPSY14] and [NS02] with the [The20] implementation of the BKZ algorithm [SE94]. The attacks take a few minutes. The longest is the black box attack on ECDSA using CM2, which takes roughly 15 minutes on a desktop computer with AMD Ryzen 9 CPU. This is due to several retries needed to get all 80 samples correct.

Additional evaluation on a public ECC dataset To better compare to prior work, we evaluate GPAM against the REASSURE ECC dataset introduced by [Chm20]. This dataset consists of roughly 6,000 unprotected traces subdivided into 255 sub-traces that are aligned to expose the corresponding cswap bit of the Montgomery ladder. Each of the sub-traces consists of 5,500 points.

The model described in [NCOS16] predicts the cswap bits with 99.6% accuracy and therefore recovers the whole key from a single trace. GPAM achieves comparable accuracy 99.57% *without the need for hypertuning*, showcasing once again its ability to generalize across datasets and use-cases.

Evaluating prior-art models on our datasets Finally, to evaluate the difference in generalization capability between GPAM and previous work, we train the LSTM model proposed in [LZC⁺21] on ECC CM0, targeting k_0 . This model achieves 91.4% accuracy but fails at attacking CM1. This demonstrates its limitations in generalizing past simple constant-time defenses to strongly protected implementations. On the other hand it proves our point that a single model can be used for multiple algorithms ([LZC⁺21] were targeting AES).

We also train a CNN model similar to [WCPB21] to evaluate convolution networks generalization potential. We had to halve the number of filters and lower the batch size to 32 to be able to run it on our longer ECC traces. This model achieved 100% accuracy on CM0 k_0 but, similarly to [LZC⁺21], failed to show significant leakage on CM1 km_0 (mean rank 118). This highlights a similar limited ability to generalize against strong defenses.

We note that those findings are consistent with our extensive internal experiments before landing on GPAM, during which we were not able to get even state-of-the-art convolutional architecture such as ConvNeXt [LMW⁺22] to achieve results as good as GPAM's.

7 Generalizing GPAM to AES

In this section, we show that GPAM generalizes across cryptographic algorithms by demonstrating its effectiveness in attacking an AES software-protected implementation, namely the publicly available ASCADv2 dataset [MS23], in an end-to-end manner without trace processing. Additionally we evaluate GPAM on two other publicly available AES datasets namely ASCADv1 [ERR⁺18] and CHES 2023 challenge to assess its ability to generalize to various AES implementations and compare its performances to previous specialized approaches.

AES attack evaluation Leakage from a single trace is often not enough to uncover the secret key in AES. In this case, the attacker may combine information from multiple traces with the same key but variable plaintext. Predictions of key byte values are then combined over a set of attack traces to a maximum likelihood score vector (their logarithms are summed). The index in this vector with the largest value is the predicted value of the key byte. More generally, *Guessing Entropy* (GE) [SMY06] is the average number of entries larger than the one corresponding to the correct value (also mean rank of the correct value

in the score vector). When the GE of all key bytes is less than one we call the attack successful.

When a sensitive value s (e.g., value of a byte of the S-BOX input) is split into two shares, bytes x, y , we may target x and y separately and then from their probabilities we compute $P[s = b] = \sum_{i=0x00}^{0xFF} P[x = b \oplus i]P[y = i]$ for any byte value $b \in \{0x00, \dots, 0xFF\}$. Analogous formulas for other types of masking (shuffling and affine masking) are derived by [MS23].

7.1 ASCADv2 dataset

The ASCADv2 dataset [MS23] comprises 800,000 power traces collected from a Cortex M4 microcontroller manufactured by ST Microelectronics (STM32F303RCT7) while it was performing AES-128 encryptions. The firmware implements affine masking and shuffling to protect the AES encryption computation from side-channel attacks. More details about the dataset can be found in [BKPT20].

Attack Scenario We replicate the “First Threat Scenario” described in [MS23]. We target the following equation from the AES S-BOX masked operation [MS23]:

$$c[i] = r_m \times Sbox[pt[p[i]] \oplus k[p[i]]] \oplus r_{out}$$

where \times and \oplus stand for multiplication and addition in the Rijndael finite field [DR01], $Sbox$ is the AES S-BOX, r_m and r_{out} are affine mask bytes, $p[i]$ is the permutation index, $pt[i]$ is the byte of the plaintext, and $k[i]$ is the byte of the AES round key.

Critically, unlike previous work [MS23], we perform an attack by using all 1,000,000 points of each trace without preprocessing, instead of using an SNR (signal-to-noise ratio analysis) analysis to use only 15,000 points (1.5%) out of the total trace. Additionally, we do not modify the GPAM architecture to perform the attack and solely rely on hyperparameter tuning to adjust the model hyperparameters to this new target.

AES attack evaluation For $c[i]$, we report the best of 7 runs (due to the influence of initialization weights, see Section 5.3.1). That is, we train a model seven times, and for each attack point, we pick the model with the highest accuracy on the validation set. We then use that to evaluate the performance on the holdout split. Our model targets the intermediate value $c[i]$, mask bytes r_m, r_{out} , and the permutation $p[i]$ (for $i = 0, \dots, 15$). Table 11 shows a comparison of ML metrics of GPAM and the results of [MS23]. We then estimate GE to also compare attack results from the acquired intermediate values. Following [MS23] we simulate a fixed key (all zero) by replacing pt by its XOR with the corresponding random key. Since r_m has very high accuracy we use it directly instead of the whole probability distribution. We sample 10,000 times to estimate the GE and need roughly 80 traces to have GE of all key bytes under 1, comparable with the 60 traces needed by [MS23] (but use heavy preprocessing).

Table 11: ASCADv2 results ($c[i]$ is the best out of 7 runs) compared with results of [MS23], measured on holdout dataset with 80,000 examples. For the sake of brevity we average results when there are multiple indexes.

Attack point	Accuracy [%]	MeanRank	MaxRank	Acc [MS23]	MeanRank [MS23]
r_m	100.00	0	0	99.2	–
r_{out}	18.25	4.78	65	21.1	–
$c[i]$ (average)	1.18	80.65	255	1.6	80
$p[i]$ (average)	95.07	0.055	4.44	88.9	–

7.2 ASCADv1 variable key

ASCADv1 the precursor of ASCADv2 is comprised of two electromagnetic emission datasets captured of an ATmega8515 micro-controller running a masked AES implementation. Following [EST⁺22] we target the dataset with variable key and compare to the SOTA attacks [LZC⁺21] and [HCM24]. For this attack we hypertune the patch size (possible values [100, 200, 400, 625, 1000, 2000]) and merge filter 1 (possible values [0, 4, 8, 16, 32, 64]) Each tuning run comprise 50 epochs of 500 steps and use a batch size of 256. The learning rate is set to 0.0005 and we use the full trace length. A total of 49 experiments run over two days with the best model having a merge filter 1 equal to 0 and a patch size of 625. This configuration was fully trained for 500 epochs and achieves a 95.94% accuracy on the third byte of S-BOX input which is the standard target as the first two bytes in the dataset are unprotected due to the mask being always zero. With 95.94% GPAM significantly outperforms both [LZC⁺21] which only achieve 6% accuracy and [HCM24] neither of which achieves a single trace attack. Full training of GPAM required twice as long as [LZC⁺21] on similar hardware but reached 90% validation accuracy halfway through.

7.3 CHES 2023 SMAesH challenge

We evaluate GPAM on the the CHES 2023 conference challenge⁵ consisting of two protected AES datasets A7 (Artix-7 FPGA) and S6 (Spartan-6 FPGA). GPAM uncovers leakage but does not achieve top results.

S6 dataset: we target bytes 1, 6, 11, and 12 as suggested by the winner. GE with 250k traces was under 15 (0.85 in one case). Other bytes did not leak enough for an attack.

A7 dataset: a single model targets $\text{msk_plaintext}[i] \oplus \text{msk_key}[i]$ and $\text{msk_plaintext}[i+16] \oplus \text{msk_key}[i+16]$ (since the XOR of these is the XOR of the original key and plaintext). These targets reach 124 mean rank. For some i one of those values did not converge and prevented the recovery of the corresponding key byte (re-training might be beneficial). Best GE at 290k traces (current winner) was 23.7 (for $i = 2$) which is still too high for an attack.

8 Conclusion

In this paper we presented GPAM, the first deep-learning architecture that can perform power side-channel analysis against multiple protected cryptographic algorithms, namely ECC and AES. We demonstrate GPAM’s ability to generalize by successfully attacking several highly protected ECC implementations without changing the model architecture, and verify its effectiveness on multiple devices. To enable reproducibility of these results, we open-source our models and datasets. This research moves us one step closer to a fully-automated side-channel attacks and leakage detection system that could be used as part of a hardware product release process and to test new countermeasures. Our results also suggest that some advanced countermeasures that are currently considered sufficient to thwart side-channels attacks can be defeated. Accordingly, there is a pressing need to devise new countermeasures that are resilient to deep-learning attacks.

References

- [AAB⁺15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur,

⁵<https://smaesh-challenge.simple-crypto.org/>

- Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [AGF23] Rabin Y. Acharya, Fatemeh Ganji, and Domenic Forte. Information theory-based evolution of neural networks for side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2023.
- [B⁺19] Elie Bursztein et al. SCAAML: Side channel attacks assisted with machine learning, 2019.
- [BCH⁺20] Shivam Bhasin, Anupam Chattopadhyay, Annelie Heuser, Dirmanto Jap, Stjepan Picek, and Ritu Ranjan. Mind the portability: A warriors guide through realistic profiled side-channel analysis. In *Network and Distributed System Security Symposium*, 2020.
- [BJSR22] Dara Bahri, Heinrich Jiang, Tal Schuster, and Afshin Rostamizadeh. Is margin all you need? an extensive empirical study of active learning on tabular data. *arXiv preprint arXiv:2210.03822*, 2022.
- [BKPT20] Ryad Benadjila, Louiza Khati, Emmanuel Prouff, and Adrian Thillard. Hardened library for AES-128 encryption/decryption on ARM Cortex M4 architecture. <https://github.com/ANSSI-FR/SecAESSTM32>, 2020.
- [BP19] Elie Bursztein and Jean-Michel Picod. A hacker guide to deep learning based side channel attacks. In DEF CON, editor, *DEF CON 27*, 2019.
- [BR23] Sonia Belaïd and Matthieu Rivain. High order side-channel security for elliptic-curve implementations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2023.
- [BVdPSY14] Naomi Benger, Joop Van de Pol, Nigel P. Smart, and Yuval Yarom. “Ooh Aah... Just a Little Bit”: a small amount of side channel can go a long way. In *Cryptographic Hardware and Embedded Systems—CHES 2014: 16th International Workshop, Busan, South Korea, September 23-26, 2014. Proceedings 16*. Springer, 2014.
- [C⁺15] François Chollet et al. Keras. <https://keras.io>, 2015.
- [Car98] Rich Caruana. *Multitask learning*. Springer, 1998.
- [CCC⁺19] Mathieu Carbone, Vincent Conin, Marie-Angela Cornelié, François Dassance, Guillaume Dufresne, Cécile Dumas, Emmanuel Prouff, and Alexandre Venelli. Deep learning to evaluate secure RSA implementations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2019.
- [CDP17] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Convolutional neural networks with data augmentation against jitter-based countermeasures: Profiling attacks without pre-processing. In *Cryptographic Hardware and Embedded Systems—CHES 2017: 19th International Conference, Taipei, Taiwan, September 25-28, 2017, Proceedings*, 2017.
- [Chm20] Łukasz Chmielewski. Reassure (h2020 731591) ECC dataset, 2020.
- [Cho21] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

- [CK14] Omar Choudary and Markus G. Kuhn. Efficient template attacks. In *Smart Card Research and Advanced Applications: 12th International Conference, CARDIS 2013, Berlin, Germany, November 27-29, 2013. Revised Selected Papers 12*, 2014.
- [Cor99] Jean-Sébastien Coron. Resistance against differential power analysis for elliptic curve cryptosystems. In *CHES*, 1999.
- [CRR03] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template attacks. In *Cryptographic Hardware and Embedded Systems*, 2003.
- [CTB⁺21] Pu-Chin Chen, Henry Tsai, Srinadh Bhojanapalli, Hyung Won Chung, Yin-Wen Chang, and Chun-Sung Ferng. A simple and effective positional encoding for transformers, 2021.
- [CZH⁺21] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training, 2021.
- [DDME⁺18] Fergus Dall, Gabrielle De Micheli, Thomas Eisenbarth, Daniel Genkin, Nadia Heninger, Ahmad Moghimi, and Yuval Yarom. Cachequote: Efficiently recovering long-term secrets of sgx epid via cache attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2018.
- [DR01] Joan Daemen and Vincent Rijmen. Reijndael: The advanced encryption standard. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 26(3):137–139, 2001.
- [ERR⁺18] Prouff Emmanuel, Strullu Remi, Benadjila Ryad, Cagli Eleonora, and Dumas Cecile. Study of deep learning techniques for side-channel analysis and introduction to ASCAD database. *CoRR*, 2018.
- [EST⁺22] Maximilian Egger, Thomas Schamberger, Lars Tebelmann, Florian Lippert, and Georg Sigl. A second look at the ascad databases. In *International Workshop on Constructive Side-Channel Analysis and Secure Design*, pages 75–99. Springer, 2022.
- [Fid15] Fido Alliance. FIDO 2.0: Key Attestation Format, 2015.
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [GJS20] Aron Gohr, Sven Jacob, and Werner Schindler. Subsampling and knowledge distillation on adversarial examples: New techniques for deep learning based side channel evaluations. Cryptology ePrint Archive, Report 2020/165, 2020. <https://eprint.iacr.org/2020/165>.
- [GLS22] Aron Gohr, Friederike Laus, and Werner Schindler. Breaking masked implementations of the clyde-cipher by means of side-channel analysis - A report on the CHES challenge side-channel contest 2020. Cryptology ePrint Archive, Report 2022/471, 2022. <https://eprint.iacr.org/2022/471>.
- [GRV17] Dahmun Goudarzi, Matthieu Rivain, and Damien Vergnaud. Lattice attacks against elliptic-curve signatures with blinded scalar multiplication. In *Selected Areas in Cryptography (SAC)*, 2017.

- [HCM24] Suvadeep Hajra, Siddhartha Chowdhury, and Debdeep Mukhopadhyay. Es-tranet: An efficient shift-invariant transformer network for side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2024(1):336–374, 2024.
- [HDLL22] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *International Conference on Machine Learning*. PMLR, 2022.
- [HGG19] Benjamin Hettwer, Stefan Gehrler, and Tim Güneysu. Deep neural network attribution methods for leakage analysis and symmetric key recovery. In *Selected Areas in Cryptography*, 2019.
- [HGS01] Nick A Howgrave-Graham and Nigel P. Smart. Lattice attacks on digital signature schemes. *Designs, Codes and Cryptography*, 2001.
- [HHGG20] Benjamin Hettwer, Tobias Horn, Stefan Gehrler, and Tim Güneysu. Encoding power traces as images for efficient side-channel analysis. In *2020 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pages 46–56, 2020.
- [HIM⁺14] Johann Heyszl, Andreas Ibing, Stefan Mangard, Fabrizio De Santis, and Georg Sigl. Clustering algorithms for non-profiled single-execution attacks on exponentiations. In *CARDIS*, 2014.
- [JB17] Kimmo Järvinen and Josep Balasch. Single-trace side-channel attacks on scalar multiplications with precomputations. In *CARDIS*, 2017.
- [JMV01] Don Johnson, Alfred Menezes, and Scott Vanstone. The elliptic curve digital signature algorithm (ECDSA). *International journal of information security*, 2001.
- [JSSS20] Jan Jancar, Vladimir Sedlacek, Petr Svenda, and Marek Sys. Minerva: The curse of ECDSA nonces: Systematic analysis of lattice attacks on noisy leakage of bit-length of ECDSA nonces. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020.
- [KJJ99] Paul Kocher, Joshua Jaffe, and Benjamin Jun. Differential power analysis. In *Crypto*, 1999.
- [KMH⁺20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- [Koc96] Paul C. Kocher. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In *Advances in Cryptology*. Springer, 1996.
- [LBM15] Liran Lerman, Gianluca Bontempi, and Olivier Markowitch. A machine learning approach against a masked AES: Reaching the limit of side-channel attacks with a learning model. *Journal of Cryptographic Engineering*, 2015.
- [LH16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [LMW⁺22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.
- [LZC⁺21] Xiangjun Lu, Chi Zhang, Pei Cao, Dawu Gu, and Haining Lu. Pay attention to raw traces: A deep learning architecture for end-to-end profiling attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2021.

- [MBC⁺20] Loïc Masure, Nicolas Belleville, Eleonora Cagli, Marie-Angela Cornélie, Damien Couroussé, Cécile Dumas, and Laurent Maingault. Deep learning side-channel analysis on large-scale traces - A case study on a polymorphic AES. Cryptology ePrint Archive, Report 2020/881, 2020. <https://eprint.iacr.org/2020/881>.
- [MOP08] Stefan Mangard, Elisabeth Oswald, and Thomas Popp. *Power analysis attacks: Revealing the secrets of smart cards*. Springer Science & Business Media, 2008.
- [MPP16] Houssein Maghrebi, Thibault Portigliatti, and Emmanuel Prouff. Breaking cryptographic implementations using deep learning techniques. In *Security, Privacy, and Applied Cryptography Engineering (SPACE)*, 2016.
- [MS23] Loïc Masure and Rémi Strullu. Side-channel analysis against ANSSI’s protected AES implementation on ARM: end-to-end attacks with multi-task learning. *Journal of Cryptographic Engineering*, 2023. <https://eprint.iacr.org/2021/592.pdf>.
- [MSEH20] Daniel Moghimi, Berk Sunar, Thomas Eisenbarth, and Nadia Heninger. TPM-Fail: TPM meets timing and lattice attacks. In *USENIX Security Symposium*, 2020.
- [NC18] Erick Nascimento and Łukasz Chmielewski. Applying horizontal clustering side-channel attacks on embedded ecc implementations. In *CARDIS*, 2018.
- [NCOS16] Erick Nascimento, Lukasz Chmielewski, David Oswald, and Peter Schwabe. Attacking embedded ECC implementations through cmov side channels. In Roberto Avanzi and Howard M. Heys, editors, *SAC 2016*, volume 10532 of *LNCS*, pages 99–119. Springer, Heidelberg, August 2016.
- [New] NewAE. NewAE Technology Inc. K82F target for CW308. <https://www.newae.com/ufo-target-pages/NAE-CW308T-K82F>.
- [NS02] Phong Q. Nguyen and Igor E. Shparlinski. The insecurity of the digital signature algorithm with partially known nonces. *Journal of Cryptology*, 2002.
- [NS19] Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*, 2019.
- [OBL⁺19] Tom O’Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Kerastuner. <https://github.com/keras-team/keras-tuner>, 2019.
- [PB19] Jean-Michel Picod and Elie Bursztein. Deep learning revolutionizing side channel cryptanalysis. DEF CON 27 <https://www.youtube.com/watch?v=QXTricqAtPk>, 2019.
- [PCBP21] Guilherme Perin, Łukasz Chmielewski, Lejla Batina, and Stjepan Picek. Keep it unsupervised: Horizontal attacks meet deep learning. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2021.
- [PCP20] Guilherme Perin, Łukasz Chmielewski, and Stjepan Picek. Strength in numbers: Improving generalization with ensembles in machine learning-based profiled side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020.

- [PP21] Guilherme Perin and Stjepan Picek. On the influence of optimizers in deep learning-based side-channel analysis. In *Selected Areas in Cryptography*, 2021.
- [PPM⁺23] Stjepan Picek, Guilherme Perin, Luca Mariot, Lichao Wu, and Lejla Batina. SoK: Deep learning-based physical side-channel analysis. *ACM Computing Surveys*, 2023.
- [QS01] Jean-Jacques Quisquater and David Samyde. Electromagnetic analysis (EMA): Measures and counter-measures for smart cards. In *Smart Card Programming and Security: International Conference on Research in Smart Cards*, 2001.
- [RBA20] Unai Rioja, Lejla Batina, and Igor Armendariz. When similarities among devices are taken for granted: Another look at portability. In *AFRICACRYPT*, 2020.
- [RIL20] Thomas Roche, Laurent Imbert, and Victor Lomné. Side-channel attacks on blinded scalar multiplications revisited. In *CARDIS*, 2020.
- [RLMI21] Thomas Roche, Victor Lomné, Camille Mutschler, and Laurent Imbert. A side journey to titan. In *USENIX Security Symposium*, 2021.
- [Rud17] Sebastian Ruder. An overview of multi-task learning in deep neural networks, 2017.
- [RZL17] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7(1):5, 2017.
- [SE94] Claus-Peter Schnorr and Martin Euchner. Lattice basis reduction: Improved practical algorithms and solving subset sum problems. *Mathematical programming*, 1994.
- [SI11] Werner Schindler and Kouichi Itoh. Exponent blinding does not always lift (partial) SPA resistance to higher-level security. In *ACNS*, 2011.
- [SMY06] Francois-Xavier Standaert, Tal G. Malkin, and Moti Yung. A unified framework for the analysis of side-channel key recovery attacks (extended version). Cryptology ePrint Archive, Report 2006/139, 2006. <https://eprint.iacr.org/2006/139>.
- [Tel] Teledyne. Teledyne LeCroy WavePro 404HD-MS. <https://teledynelecroy.com/oscilloscope/wavepro-hd-oscilloscope/wavepro-404hd-ms>.
- [The20] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 9.0)*, 2020. <https://www.sagemath.org>.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.
- [WAGP20] Lennert Wouters, Victor Arribas, Benedikt Gierlichs, and Bart Preneel. Revisiting a methodology for efficient CNN architectures in profiling attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020.
- [WCPB21] Léo Weissbart, Łukasz Chmielewski, Stjepan Picek, and Lejla Batina. Systematic side-channel analysis of Curve25519 with machine learning. Cryptology ePrint Archive, Report 2021/944, 2021. <https://eprint.iacr.org/2021/944>.

- [WHJ⁺21] Yoo-Seung Won, Xiaolu Hou, Dirmanto Jap, Jakub Breier, and Shivam Bhasin. Back to the basics: Seamless integration of side-channel pre-processing in deep neural networks. *IEEE Transactions on Information Forensics and Security*, 2021.
- [WP20] Lichao Wu and Stjepan Picek. Remove some noise: On pre-processing of side-channel measurements with autoencoders. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020.
- [WPB19] Leo Weissbart, Stjepan Picek, and Lejla Batina. One trace is all it takes: Machine learning-based side-channel attack on EdDSA. In *Security, Privacy, and Applied Cryptography Engineering: 9th International Conference, SPACE 2019, Gandhinagar, India, December 3–7, 2019, Proceedings 9*, pages 86–105. Springer, 2019.
- [WPP22] Lichao Wu, Guilherme Perin, and Stjepan Picek. I choose you: Automated hyperparameter tuning for deep learning-based side-channel analysis. *IEEE Transactions on Emerging Topics in Computing*, 2022.
- [XTG⁺20] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.
- [ZBHV20] Gabriel Zaid, Lilian Bossuet, Amaury Habrard, and Alexandre Venelli. Methodology for efficient CNN architectures in profiling attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020.
- [ZBHV21] Gabriel Zaid, Lilian Bossuet, Amaury Habrard, and Alexandre Venelli. Efficiency through diversity in ensemble models applied to side-channel attacks:—a case study on public-key algorithms—. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2021.
- [ZS20] Yuanyuan Zhou and François-Xavier Standaert. Deep learning mitigates but does not annihilate the need of aligned traces and a generalized ResNet model for side-channel attacks. *Journal of Cryptographic Engineering*, 2020.
- [ZSX⁺20] Fan Zhang, Bin Shao, Guorui Xu, Bolin Yang, Ziqi Yang, Zhan Qin, and Kui Ren. From homogeneous to heterogeneous: Leveraging deep learning based power analysis across devices. In *ACM/IEEE Design Automation Conference (DAC)*, 2020.