Statistical evaluation of PUF implementation techniques as applied to quantum confinement semiconductors

Thomas Patrick McGrath

Department of Physics

Lancaster University

April 2021

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Preface

This thesis is the result of work which I performed at Lancaster University, between April 2017 and March 2021. Except where otherwise stated the contents of this thesis are the results of my own work, and is not the same as any others I have already submitted, or in the process of submitting, for any degree at any university or institution. The word count on this thesis does not exceed the maximum limit of 80,000 words.

T. P. McGrath

Acknowledgements

Words cannot capture the depth and sincerity of my gratitude towards those who have helped me, and shown me kindness, throughout this project. Know that the following, therefore, in no way gives justice to the profundity of the sentiment it attempts to illustrate.

First, I would like to express the extremity of my gratitude towards my supervisor, Robert Young, for his support and guidance, both academic and otherwise, throughout this project. I cannot conceive of a more ideal supervisor, nor a better positive influence and role model. I owe you a lot.

Next, I wish to recognise the members of the research group, both new and old, that have made my time involved therein so meaningful. James, Elliot, Kieran, Chris, Ramon, Povilas, Utz, Ethem, Somak, Yameng and Yasir have my sincerest thanks for many memories and imparted knowledge.

This respect and indebtedness continue to the many collaborators and helpful influences outside of the research group that has made the completion of the project attainable. This list is large, and giant is each shoulder I have stood on, extending in number well beyond the few especially named in the contribution section overleaf.

Finally, I would like to express my warmest appreciation for my family and many friends that were by my side during the time of this project. While not as direct, the contributions of these wonderful people are just as irreplaceable, and worthy of dedication, as any other here.

I would, however, here also like to acknowledge (rather than express any form of gratitude for) the COVID-19 pandemic, without which this work would be very different.

Thank you.

Contributions

The resonant tunnelling diodes characterised in this study were designed and fabricated by J. Sexton and M. Missous at the School of EEE, University of Manchester. The electronic measurements of the diodes were taken with the aid of S. Dean and C. Barthelmes at Lancaster University. No further direct contributions to this work exist to declare.

Summary

Physically unclonable functions, or PUFs, present a means to securely identify objects, both implicit and attached, alongside several uses in conventional secure communication techniques. Many types of PUF based on varying sources of fingerprint entropy have been suggested, and the higher-level theoretical properties and implications of this primitive have been extensively discussed.

However, each different prospective implementation of PUF typically approaches the practical considerations for the conversion from a unique entropy source to ultimate PUF implementation anew. These studies typically treat the intermediate processing schema, such as response binning, solely as a means to an end rather than a subject of explicit discussion and evaluation. As such, there exist few studies into developing a general framework for the optimisation and simulation of the important elements that lie between the measurement of the particular entropy source and the evaluation of the final device as a whole.

This thesis seeks to outline and validate a generalised schema for the conversion of entropy source to final results, presenting the fundamental design elements and figures of merit for the process at every stage where applicable. Further to this, each stage of the process is expressed analytically, allowing the direct derivation of the ultimate figures of merit based on the measurement outcomes of the initial source of entropy. To validate, this process is applied towards the resonant tunnelling diode (RTD) as the prospective entropic unit cell. This type of semiconductor device has several properties that make it an interesting candidate upon which to base a PUF, and this work additionally seeks to outline these benefits and enumerate the general comparative figures of merit for a PUF derived therefrom.

List of Publications

N. Abdelazim, J. Fong, **T. McGrath**, C. Woodhead, F. Al-Saymari, I.E. Bagci, A. Jones, X. Wang, R. Young, “Hotspot Generation for Unique Identification with Nanomaterials”, Scientific Reports **11**, 1528, (2021).

T. McGrath, I.E. Bagci, Z. Wang, U. Roedig, R.J. Young, “A PUF Taxonomy”, Applied Physics Reviews **6**, 1, 011303, (2019).

I.E. Bagci, **T. McGrath**, C. Barthelmes, S. Dean, R. Bernardo Gavito, R.J. Young, U. Roedig, “Resonant-Tunnelling Diodes as PUF Building Blocks”, IEEE Transactions on Emerging Topics in Computing, 10.1109/TETC.2019.2893040 (2019).

R.J. Young, I.E. Bagci, **T. McGrath**, U. Roedig, R. Bernardo Gavito, “System and method for generating entropy values”, GB2575040A, (2018).

R. Bernardo Gavito, I.E. Bagci, J. Roberts, J. Sexton, B. Astbury, H. Shokeir, **T. McGrath**, Y. Noori, C. Woodhead, M. Missous, U. Roedig, R.J. Young, “Extracting random numbers from quantum tunnelling through a single diode”, Scientific Reports **7**, 17879, (2017).

R. Bernardo Gavito, F. J. Urbanos, J. Roberts, J. Sexton, B. Astbury, H. Shokeir, **T. McGrath**, Y. Noori, C. Stephen Woodhead, M. Missous, U. Roedig, R.J. Young, “N-state random switching based on quantum tunnelling”, Proceedings of SPIE, 10354 (2017).

Contents

Chapter 1: Introduction

In a world that is becoming increasingly interconnected and digitally abstracted, dependency ever increases on systemic schemata of trust and validation. One valuable solution in establishing secure communication and authentication comes in the form of physically unclonable functions, or PUFs. These are devices that convert some element of hard-to-recreate physical entropy in a system into a digital key or signature. Performing this process associates the object hosting the physical entropy with a unique fingerprint, which can later be checked to verify the identity of the physical object. However, countless physical systems can be employed as PUFs, and it is not readily apparent which systems provide the best foundation to build upon. This work attempts to determine and example an analytical schema to translate the measured properties of a physical system into resultant PUF metrics. This will aid both the determination of what PUF entropy concepts are most viable and make apparent the direct consequences of variations to a particular PUF based on the adjustment of the means of physical entropy measurement, allowing for the impact of improvements to this entropy measurement to be simulated without fully and physically manifesting the more general systemic elements of the PUF.

This work starts with a section more thoroughly introducing the concept of PUFs and the basic physics of resonant tunnelling diodes, or RTDs. These diodes are electronic devices that exhibit an unusual non-monotonicity in current with increasing voltage as a result of quantum effects. This N-shaped electronic characteristic not only provides an easily definable feature from which to extract a value, but its apex position is determinant on the unique configuration of a relatively small number of atoms magnified by a quantum well, making the recreation of such a device extraordinarily hard. As such, these elements can be considered very pertinent candidates for the source of physical entropy in a PUF, being evaluated and optimised as the second part of this work, in parallel and to help contextualise the generic elements here included. The chapter following this background discusses the parameters and methods by which these RTDs were produced and measured to provide the data sets from which the analysis found in later sections is derived.

After introducing the theory and experimental techniques, this work first focuses on the means by which an associated initial analogue value can be extracted from each RTD. This is done by examining a range of value extraction techniques applied to the diodes and introducing figures of cost and merit to compare the resultant distributions. Once a selection of optimal extraction techniques is determined, the next stage, and section in this work, can be considered as the binarisation of these analogue values. This converts each PUF entropy element into a (typically single) bit to be concatenated into a longer digital signature. This work introduces and discusses the merits of four binarisation techniques, which can be considered as most fundamental, and derives relationships between the analogue-measurement metrics of the previous section (based primarily around standard deviations) and the more conventional, higher-level PUF metrics (most notably error rate per bit). These techniques are applied to the RTD values extracted from the section prior, and the optimal of these configurations are taken forward into the final step. This final step in the PUF simulation process, and penultimate section, discusses the systemic considerations of the PUF implementation, first looking into which error reduction techniques do not leak auxiliary data that may impede predictability, and finding the optimal configuration of these. This section also outlines the process of concatenating single PUF bits into full responses, splitting the error rate into that of false positive and false negative, and briefly discussing the security standards to which a PUF would be held in these domains. The employment of error reduction techniques (decreasing error rate at the cost of measurement requirements) towards a certain security level standard allows the previously incomparable frontrunner configurations of extraction and binarisation techniques to be finally directly related. In this way, these potential configurations can be reduced to an ultimate optimal for a variety of use cases, and the expectant metrics for a PUF derived from RTDs as measured in this work can be announced, along with a complete guide to performing the same analysis with any other source of physical entropy. The final section of this work presents a conclusion that brings together the results from each section and discusses further work that can be done in expanding the scope of this study.

Chapter 2: Theory

To evince the quality of the PUFs developed in this work, one must first seek to understand an amount about the nature PUFs and the conventions of the field in general. Equally, to optimise the positive qualities of the implementations considered here, a basic understanding of the physics underpinning the RTD is of value. As such, the first part of this section introduces the central categorisations and figures of merit of the PUF. The section continues by going on to outline and classify several examples of physical systems employed as PUFs, and their relative merits and applications. Finally, the section discusses the physical principles of RTD operation and concludes by describing how this physical operation would be employed towards producing a PUF.

2.1. Physically Unclonable Functions (PUFs)

The Physically Unclonable Function, or PUF, is a hardware security fundamental with many applications, particularly in the fields of authentication and secure communication [1]. It is an entity that translates a certain input (or challenge) into a unique output (or response) in a manner that is not physically cloneable. This is to say that the set of challenge-response pairs (CRPs) are different in each PUF that is manufactured (and is therefore unique) and that no instance of the PUF can be deliberately manufactured to have the same CRP set as another (and so are unclonable). In this way, these devices can be considered as a physically embodied cryptographic hashing function, with a unique mapping for each device that is impossible to recreate. These properties allow the PUF to act as a fingerprint, verifying the unique identity of the PUF itself, and by extension any object or entity that the PUF is embedded within or attached to. In a typical authentication operation, when the PUF is manufactured the device is initially enrolled, where responses are recorded from one or more input challenges. Once these pairs are taken and the device is distributed, it can be known that the only entity that can correctly respond to those recorded challenges, if later reapplied, is the device itself. To achieve this, a PUF must be stable over time and repeated evaluations, and be easy to (uniquely) evaluate but difficult to replicate or predict. These properties are on a sliding scale and can be used to evaluate the merits of a certain PUF implementation. At a more fundamental level, however, PUF implementations can be categorised by their implementation strength, in the following section.

2.1.1. Strong and Weak PUFs

One of the important differentiations between different types, or implementations, of PUF, is their strength. This classification consists of two categories – strong PUFs and weak PUFs [2]. This strength property relates to the number of CRPs that a single device can support or, more directly, the rate at which the CRP set grows as the device increases in size (or number of constituent physical elements grow). Weak PUFs typically scale at a linear or low-order polynomial rate with increasing device size while strong PUFs scale at higher-order polynomial or exponential rates, leading to a very significantly lower or higher number of CRPs for a device of reasonable size.

For a weak PUF this means that while the physical device itself cannot be copied, an attacker with physical access to the PUF may be able to copy the full set of CRPs from the device. This would allow the attacker to convincingly respond to an authentication query, even after the device leaves their possession. Weak PUFs can therefore be used for the generation and storage of secure keys and the authentication of itself and attached objects, with the caveat that the authenticating party must be present and not, for instance, validating remotely. This is needed to verify that the response is being read from the device itself and not instead from an emulation of the PUF using the full CRP set retrieved from a PUF at a previous encounter.

Strong PUFs, on the other hand, are much more versatile. Their much larger challenge-response space prevents an attacker from acquiring or storing more than a negligible amount of the total CRP set should they have physical access to the device. This means that provided the CRPs stored and employed by the authenticating party are random, only the holder of the PUF at the exact time of authentication will be able to produce the appropriate response, facilitating for instance remote authentication. Furthermore, with such a large CRP set it may be possible to use each individual challenge-response pair only once. This operation has several useful implications. One example is that this system would prevent man-in-the-middle, or replay, attacks, where an attacker eavesdrops on a PUF authentication interaction to be able to convincingly repeat the interaction instead of the PUF at a later time. In addition to aiding secure authentication, this redundant operation allows for secret communication between parties using the PUF as a cryptographic one-time-pad. This means of communication is completely secure without access to the PUF itself.

2.1.2. Figures of Merit

To validate and compare different PUF variations and implementations, a scheme to evaluate efficacy should be outlined. This scheme examines how well any given device fulfils the required properties of a PUF and can be considered to have three sections, in order of importance. First, the PUF must be evaluated on the qualitative properties that indicate the PUFs fundamental macroscopic feasibility, such as how easy the PUF is to manufacture, evaluate and exactly how difficult a device would be to physically clone. Next, the output of that PUF implementation must be evaluated from a quantitative statistical perspective, for comparison and ensuring that the rate of false-positive or false-negative evaluations is negligible (and so each manufactured PUF can be confidently identified). This statistical evaluation also aims to ensure that there is no underlying predictable pattern in the responses that can devalue the security of the PUF. Finally, the physical specifications of merit should be evaluated, again for comparison and in checking for undue limitation. These specifications are primarily space (including bit density), time (throughput) and power requirements for any PUF implementation to operate. This last set of merits does not impact the validity of the PUF, but can impact the overall viability of a certain PUF for a certain purpose, and provides a means to contrast ostensibly similar PUFs.

2.1.2.1. Qualitative Appraisal

The three main properties of a PUF to qualitatively evaluate are the constructability, evaluability and clonability of a prospective device. However, here is considered any additional meritorious features of the PUF along similar qualitative lines. An example of this would be any enhanced tamper evidence, voiding or protection such as that which is an inherent feature of the coating style of PUFs [3, 4]. This PUF conception layers the source of entropy around the outside of the measurement and evaluation apparatus as a coating, such that any attempts to access measurements or circuitry inside directly alters the state of the entropy and invalidates further measurements of the device. A second example of a qualitative extra feature is the ability to reset or refresh sets of challenge-response pairs after manufacture, as can be found in the family known as reconfigurable PUFs (rPUFs) [5]. These typically utilise non-volatile memory in a way as to repeatedly change the state of a given unit cell randomly, to allow for extensions such as securely reassigning devices and enrolling new CRPs after old ones have been used.

2.1.2.1.1. Constructability

The first area in which a PUF can be qualitatively evaluated is its constructability. This is simply an examination of how easy or feasible the PUF is to manufacture from a qualitative perspective. PUFs that are easy to manufacture can be considered more meritorious, all things being equal, than PUFs that introduce more additional cost or hardware footprint. While this is often a value judgement, PUFs for object authentication can be split into two groups for the property of ease of construction. This is a delineation based on whether the source of randomness can be considered implicit or explicit [6]. PUFs can be considered as consisting of an entropy source and a means of evaluation, applied to an entity that is to be the object of authentication. If this entropy source does not require any additional fabrication steps, beyond the authentication object's manufacture process, to introduce, it can be considered as a source of implicit randomness. Conversely, if additional steps are required to introduce the required entropy source then the PUF is said to consist of explicit randomness. As well as reducing the cost and complexity of manufacture, PUFs derived from implicit randomness are typically considered more secure than explicit randomness PUFs. This is because with no additional independent steps to produce the entropy it is much harder for an adversary, even one with influence over the manufacturing process as a whole, to manipulate the entropy of the PUF at the point of manufacture without having notable effects on the authentication object itself.

2.1.2.1.2. Evaluability

The next area in which a prospective PUF must succeed is in its evaluability, this being whether or not the prospective entropy source of a certain PUF concept lends itself to reasonable evaluation. This domain considers the qualitative elements of the evaluability of the PUF, or how readable the PUF is in the abstract, in contrast to the more specific elements of bit density, throughput and power consumption that allow quantitative comparison between PUF concepts (once validated in general terms) later in this section. Again, all things being equal a PUF that is inherently easier to evaluate outshines a PUF that makes it harder to extract the response from the entropy source. Also, while the ease of evaluation for each given implementation is on a gradient, there exist two distinct groupings within this category – intrinsic and extrinsic evaluation [7]. An intrinsic PUF is a PUF whose entropy arises implicitly and can internally evaluate, whereas an extrinsic PUF has means of evaluation external to the PUF itself. In other words, an intrinsic PUF has the means of probing the characteristic entropy of the system in a manner embedded within, or intrinsic to, the device itself. Since this evaluation typically involves measuring the entropy and converting it to an electronic signal for communication, generally only implicit-entropy PUFs based on authenticating electronic devices can be considered as intrinsic. In general, intrinsically evaluating PUFs tend to be more accurate, easier to use and less prone to security attacks. These internally evaluating PUFs have the attractive feature that further processing on the entropy response, for instance hashing or asymmetric key cryptography protocols, can be performed in situ without exposing the raw entropy response to external measurement. In an extrinsic PUF, the raw entropy source must be exposed for external measurement, and so cannot undertake any steps of obfuscation against an attacker directly reading the measurement directly from the source.

2.1.2.1.3. Physical Clonability

Most fundamental to the concept of the PUF is the idea of the lack of physical clonability. That is to say that for any manufactured PUF it should be impossible to physically reproduce its characteristic output in another device. While we can say that copying a PUF should be impossible, it is worth noting that there is always the chance that two PUFs coincidentally have the same output response, proportional to the total number of CRPs the PUF supports. Assuming ideal statistical properties (where each permutation of CRPs is equally likely) this is equal to the reciprocal of the total number of possible permutations. However, an appraisal of the physical clonability of a PUF examines the case where an interested party attempts to deliberately recreate the challenge-response pairs of one PUF in another, and so skew the chance of the two PUFs being the same beyond what would be statistically anticipated. In other words, clonability is the measure of how well the uniqueness of a PUF can be enforced by design. The level of clonability of a PUF exists over many dimensions depending on the particular situation of a prospective attacker. A highly unclonable PUF should be resistant to recreation even if the attacker has access to the highest quality tools, including the equipment that was used to create the original PUF in the first place. It must not be possible to copy a PUF even with complete physical access to it, or by modifying another PUF to give the same response. Some PUF designs, such as the optical and electronic quantum PUFs, maintain that they give such a unique response signature that it is possible to verify that the source of entropy for the PUF is of the same (unclonable) material [8, 9]. This helps resist the use of a PUF using a completely different and more controllable material system to simulate the PUF response and adds an additional layer to the device's clonability (simulation attack resistance).

2.1.2.2. Statistical Appraisal

After examining a prospective PUF design for qualitative feasibility and merit, an evaluation of the statistical properties of the PUF should be undertaken. A PUF must be of adequate security from a statistical perspective to be viable and if, for instance, achieving a reasonable level of entropy for security prohibitively increases the footprint of the device then a certain PUF concept cannot be considered as a candidate for employment. To analyse PUFs statistically, the set of responses from the PUF can be considered as consisting of three independent dimensions. The first of these is the variation of the response bits within the same PUF, known as the uniformity of the response. The second dimension is the variation of the same response upon multiple evaluations, known as reliability. The final dimension is the variation of the CRP at the same location on a PUF for different manufactured instances of that PUF, known as the uniqueness of the PUF. Different CRPs on the same

PUF or across different PUFs should be as close to independent as possible, whereas the measured output of any given CRP should be the same across repetitions.

There also exist metrics for determining, for instance, the independence, uniformity, and reliability of different sets of multi-bit challenge-response pairs in the same PUF. Examples of this would be the diffuseness and steadiness in certain works [10], and bit aliasing in others [11]. These metrics are only applicable for cases where the response in a single challenge-response pair is more than a single bit in length. In this work, however, the RTD based PUFs that are simulated are all designed to output a single standalone bit per evaluation. Keys of cryptographically secure length are built up from concatenating these single-bit response units in a random order, employing the entire PUF for the weak PUFs simulated in this work. This means that no matter what is considered the response of the PUF (be it the single bit or the concatenated key), these inter-ID and intra-ID metrics cannot apply. Looking at randomised single-bit CRPs means that the uniformity and reliability of the CRP are equal to the uniformity and reliability of the whole PUF and so the uniformity, reliability and uniqueness between CRPs are the same as for the PUF as a whole.

Ultimately, uniqueness and reliability act to influence the identifiability of the PUF, this being the confidence by which one can determine a PUF measurement to be accurate. The uniqueness and the uniformity form the basic predictability, or advantage an attacker can gain from a disparity in the 0/1 likelihood in predicting further responses.

2.1.2.2.1. Uniformity/Bias

The uniformity, otherwise known as the bias, of a PUF, represents the variation of response bits within the same manufactured instance of a PUF. In other words, this property is a test to determine the probability of a ‘0’ or a ‘1’ occurring at any point along the PUF response set. To be secure, the proportion of ‘0’ to ‘1’ bits in a PUF should be equal, and therefore of equal (50% each) probability of occurring, leading to equal representation. This would mean that an attacker has no more than a 50% chance of guessing the next response bit correctly, and cannot gain a statistical advantage. This value can be calculated by examining the fractional Hamming distance of the bits in the device, and so can be described as the intra-Hamming-distance of the PUF, or the inter-Hamming-distance of a response bit with respect to a single PUF instance. Hamming distance is the count of the number of positions where values differ across two compared strings (here bit strings) of equal length. Fractional hamming distance divides this count by string length, to derive a value that is independent of string length (a length here equivalent to response length of a full PUF instance). The unadjusted uniformity, here called uniformity prime or bias, of a PUF instance i consisting of P total response bits is equal to the expectation value and can be described as:

$$(Uniformity')_i = (Bias)_i = \frac{2}{P(P-1)} \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} D_H(r_{pi}, r_{qi}) \; = \frac{1}{P} \sum_{p=1}^{P} r_{pi} \tag{1}$$

Where r_{pi} (or r_{qi}) represents the 0/1 bit states (string length 1) of a response bit at position p (or q) of PUF instance i, as the 'true' or most probable state of the response as relating to the measurement reliability of a given bit response in section 2.1.2.2.3. $D_H(x, y)$ represents the hamming distance between strings x and y, and P represents the total number of response positions (or total bit length per instance) of the PUF. This measure for uniformity can be expressed fractionally or as a percentage and is typically averaged over multiple PUF instances to build up a picture of the PUF system as a whole. As the ideal value for this property is 50%, with deleterious deviation on either side, it must be transformed before being included in averaging over instances. One can imagine two PUF instances where one instance has a surplus of '1' states and the other the same for '0's, resulting in uniformities of 25% and 75% respectively. Averaging these would result in the ideal 50% as if both instances were ideal. To remedy this, we can introduce a second property, with an ideal value of 1 and deviations only below, to result in a comparative score out of 100%. This processed uniformity is described in equation (2) below.

$$(Uniformity)_i = 1 - 2|(Uniformity')_i - \; 0.5| \tag{2}$$

An alternative transformation that allows consistent comparison is taking a processed uniformity as $0.5 + |(Uniformity')_i - \; 0.5|$, or 1 minus half of the above. This formulation is equal to the probability of the most probable bit for each case, is similar to the error/reliability transformation for averaging, and is used for the calculation of min-entropy later in this section. In this work, the unmodified (50% centred) uniformity, described as bias, is used as the metric for comparison (with some rare usage of percentage deviation from this 50% ideal bias as well). This is because the simulations in this work can be considered as being of a single instance, and since there would be no periodic correlation in simulated element position the information as to whether the bias tends towards zero or one is more useful. Assuming all PUFs have equal response bit-length, this modified value can be concisely expressed over the set of PUFs in the data set by calculating the Hamming distances of the response sets over instances using the following equation:

$$(Uniformity) = \frac{2}{LP(P-1)} \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} D_H(R_p, R_q) \tag{3}$$

Where L and P equal the total number of PUF instances and PUF response length respectively, and where $D_H(R_p, R_q)$ represents the Hamming distance between the bit vectors R_p and R_q, consisting of the response bit at position p or q respectively compiled for each instance. Note that the bit vectors

R_p and R_q represent the bit strings that consist of all response bits at certain position p and q over a collection of instances (and as such has a bit length typically greater than 1). This is in contrast to the earlier, lower case r_{pi} and r_{qi} that consist of a single bit (or bit string of length 1) for a certain response state at position p or q and specific instance i. This uniformity measurement is equivalent to the transformed single-bit and bit-averaging techniques from equation (2), as:

$$\begin{aligned}(Uniformity) &= 1-\frac{2}{L}\sum_{i=1}^{L}\left(\left|\frac{2}{P(P-1)}\sum_{p=1}^{P-1}\sum_{q=p+1}^{P} D_H\left(r_{pi},r_{qi}\right)-\ 0.5\right|\right)\\ &= 1-\frac{2}{L}\sum_{i=1}^{L}\left(\left|\frac{1}{P}\sum_{p=1}^{P} r_{pi}-\ 0.5\right|\right)\end{aligned} \tag{4}$$

With notation as before, where P represents the total number of response positions (or total bit length per instance) of the PUF and L represents the total number of instances of the PUF. Also, as before, p and q represent index values for position while i and j represent index values for instance. Finally, D_H represents the function of the hamming distance, and r represents the bit response at the subscripted index values.

Another meaningful interpretation of this property is the measure of min-entropy. This is a measure of, for each bit of measured response, the highest number of ideal, unbiased entropy bits that can be assuredly derived. As an example, a PUF that has a 75% chance of a 1 rather than a 0 in its response has a min-entropy yield of around 0.42 unbiased bits (calculated from (5)). To derive an ideal PUF response of 256 bits from this min-entropy level one would need a total of 610 bits in the response set. In the ideal case, a PUF with a 50% chance of '0' or '1' either way would yield one bit of entropy per bit of response. Min-entropy can be calculated with the following equation:

$$H_{min}(X) = \ -log_2 P_{max} \tag{5}$$

Where $H_{min}(X)$ represents the min-entropy of random variable X, and P_{max} represents the probability of the most likely outcome of that variable (equal to $0.5 + |(Bias) - 0.5|$ for a single instance).

Studies into min-entropy bring up the important point of post-processing to enhance uniformity. The independent, ideally random response set of a PUF can be considered very similar to a random number generator (RNG) output and, as is very common with RNGs, can be debiased at the cost of throughput (or, here, the total yield of responses). For instance, XORing adjacent pairs of response bits would reduce the bias by $2\beta^2$, (where β is the bias or Hamming weight from the ideal value) at the cost of halving the total number of challenge-response pairs supported by the PUF. A more effective but complex example would be conventional hashing algorithms, such as passing 256-bit blocks of responses through the SHA-256 hashing function to induce uniformity. Care must be taken to ensure that any biasing of one state over another before this processing remains hidden, as this information could still influence the predictability of the processed responses. Additionally, in PUFs with CRPs to spare, as is the case for strong PUFs, it can be argued that one can make the output much more uniform at very little practical cost using these techniques. As a constraint, it is worth noting here that variable yield debiasing techniques such as Von Neumann extractors, while maximally efficient in converting to unbiased min-entropy bits, would be hard to implement on a PUF with fixed and limited challenge and response allocations. In summary, while the vital property of uniformity can be enhanced at the cost of response bits (and thus device footprint), the amount of uniformity before processing still dictates at what cost, if feasible, this debiasing must come by. Uniformity by itself therefore still provides a very important insight into the statistical merit and viability of a given PUF implementation. For this work, the min-entropy yield will define the relationship between the number of entropy elements and the number of true entropy bits. This assumes completely ideal debiasing and can be improved for more realism in future works.

While evaluation of this property of uniformity helps to rule out basic dependencies, it is ultimately assuming that each bit of response is independent. An example of dependency would be when the three bits ‘110’ always follow the bits ‘001’, and vice versa, with all else being independent and of equal likelihood. Here a Hamming distance study would still find equal weighting and so a uniform output, but an attacker could successfully predict around 25% of the CRP set with the right information. Resistance to this style of attack, both deriving from observation or more complex machine learning attacks [12, 13], should be accounted for in a final device.

2.1.2.2.2. Uniqueness

The next fundamental statistic for the merits of a PUF is the uniqueness, tied to the false positive rate, of each instance. This property is a test to ensure that each PUF instance is unique and that there is no correlation between the responses of one PUF compared to another. This can be imagined as the same concept as the PUF uniformity, except checking for correlations over the dimension of PUF instances for a given challenge-response pair, rather than over the space, or response position dimension, of a single PUF. In the same way, the ideal outcome for security is that for each single bit response location on a given PUF, the probability of the same location on a different PUF having the same response is 50%. This means that an attacker with access to one PUF cannot measure a certain location on it to gain an advantage in guessing the bit state of the same location on another. This helps maintain an unpredictable set of responses for each PUF. Additionally, each PUF being maximally distinct contributes to how confidently an authenticator can confirm each PUF's validity, as it reduces the likelihood of two PUFs being similar as compared to the variation in their measurement. The uniqueness between instances of a PUF implementation therefore contributes to both the identifiability and predictability of the PUF itself. The uniqueness of a PUF can be calculated in the same manner as the uniformity, as with equations (1) to (4), but with the response position (P/p) and instance position (L/i) dimensions and index variables swapped. As such the ultimate figure of merit for the uniqueness, taken between the full range of 0 and 1 in the same way as uniformity, can be calculated using the following equations:

$$\begin{aligned}(Uniformity) &= \frac{2}{PL(L-1)}\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} D_H(R_i, R_j) \\ &= 1 - \frac{2}{P}\sum_{p=1}^{P}\left(\left|\frac{2}{L(L-1)}\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} D_H\left(r_{pi}, r_{pj}\right) - 0.5\right|\right) \qquad (6) \\ &= 1 - \frac{2}{P}\sum_{p=1}^{P}\left(\left|\frac{1}{L}\sum_{i=1}^{L} r_{pi} - 0.5\right|\right)\end{aligned}$$

Where L and P equal the total number of PUF instances and PUF response length respectively, where i and j act as index variables for a certain PUF instance, and p references the index variable for the response position within an instance. D_H represents the function of the Hamming distance between two bits or bit vectors, where R_i and R_j represent the bit vectors of the full response set of PUF instance i and j respectively, and r represents the bit response at the subscripted index values. It is worth noting here that in this work large collections of independent entropy elements are simulated to derive the merits of a resultant PUF. As such, there is no delineation between collections of entropy elements as there would be if these elements were spread across multiple physical devices. This means that for this work there are not inherently multiple instances of PUF to contrast, and so the uniqueness figure of merit described here can be considered equal to the uniformity, with a total number of instance groupings of $L = 1$ (with a large total number of single-instance response bit positions P) chosen for simplicity.

Unlike the reliability and uniformity of a PUF, the uniqueness property is not readily enhanced by additional processing. However, as the ideal state of this unique, inter-device, property is the same as with the uniformity, similar techniques can be employed. Both properties can be seen as the output of a random number generator, only over different dimensions and so the uniqueness debiasing can be considered analogous to the uniformity debiasing. Therefore, one possible way to enhance the overall uniqueness of a PUF system would be to produce two independent and separate PUF units on the authentication substrate and perform an XOR operation at each location to find the final response at this location. This, although very atypical, would be the equivalent of XORing two adjacent bits in the uniformity enhancement case, leading to the same positive effect, with the same cost of bits.

Furthering the parallels, it is again worth noting that an even representation of either bit state does not necessarily mean an unpredictable response. In this case, you could find a set of PUFs which, for a given response location, always alternate between '0' and '1', depending on if the manufacturing order and thus the evenness or oddness of some form of serial identifier. This would give a perfect average Hamming distance for that location, but still be highly predictable. The uniqueness property therefore assumes no underlying pattern in the procession of bits at the same location on different PUFs. As with uniformity, the correlation between inter- or intra- device response position and a response bit state, found either through observation or through machine learning attacks, should be protected against in a final device.

2.1.2.2.3. Reliability/Error Rate

The final statistical figure of merit, this being the reliability (or error rate) of a PUF, represents the variation in the resultant bits of repeated measurements of the same response of the PUF. In other words, this property is a test to determine the likelihood that a certain PUF response is measured as the same each time it is evaluated and matches, for instance, the response taken as correct in an authentication database. To be reliable, the chance that the measured PUF response is the same as the previous measurements of the same response is as high as possible, ideally at 1 (or 100%). Alternatively stated, it is optimal for the error rate to be 0% or as low as possible, where the error rate represents the rate of incorrect reading, as the compliment to reliability. This means that the authenticator can be confident that the measured response at any given time is accurate and that any variation from what is expected arises from the PUF not being the expected device rather than uncertainty in measurement. This property is very important to ensure that PUF instances are identifiable, where accommodations to an unreliable measurement accuracy can be deleterious to security or throughput. Like the previous statistics, this property is typically expressed in terms of fractional Hamming distance, also known as the intra-Hamming-distance of the response location. The reliability scoring of a response with response position p and instance i, measured M times can be described as:

$$(Reliability) = 1 - (Error\ rate) \tag{7}$$

$$(Error\ rate)_{pi} = \frac{2}{M(M-1)} \sum_{m=1}^{M-1} \sum_{n=m+1}^{M} D_H(r_{pim}, r_{pin}) = 0.5 - \left| 0.5 - \frac{1}{M} \sum_{m=1}^{M} r_{pim} \right| \tag{8}$$

Where $D_H(r_{pim}, r_{pin})$ represents the Hamming distance between the measured bits r_{pim} and r_{pin}, which in turn are the bit states of measurement index m and n of response position p and instance i. This is typically expressed as a probability or percentage, and is typically averaged over multiple challenge-response pairs and multiple PUF instances to build up a picture of the reliability over the whole system:

$$\begin{aligned}(Error\ rate) &= \frac{2}{PLM(M-1)}\sum_{p=1}^{P}\sum_{i=1}^{L}\sum_{m=1}^{M-1}\sum_{n=m+1}^{M} D_H(r_{pim}, r_{pin}) \\ &= \frac{1}{PL}\sum_{p=1}^{P}\sum_{i=1}^{L}\left(0.5 - \left|0.5 - \frac{1}{M}\sum_{m=1}^{M} r_{pim}\right|\right)\end{aligned} \tag{9}$$

Where P represents the number of response bits on a single PUF, L represents the total number of PUF instances, and where p and i represent the index values across these dimensions respectively. This work does not attempt to simulate separate devices or any periodic correlation that could constitute different instances of a physical device, and such for the studies herein we take $L = 1$. Unlike the dimensions across different PUFs or the same PUF space for uniqueness and uniformity respectively, where measurements are ideally independent, reliability measurements are examining repeated evaluation of the same CRP of the same PUF and are dependant, ideally the same. As such, contrasting to the overall uniformity being improved via debiasing methods like post-processing, final reliability can be systemically improved via majority voting, fuzzy extraction, or other error correcting codes. Here, majority voting just means taking repeated measurements for each challenge-response pair and taking the most common bit state as the actual. This improves reliability at the cost of PUF evaluation time. Care must also be taken to ensure that the error rate or reading of values do not shift with reasonable changes of conditions, such as temperature, inducing a shift in error rate for all further measurements. The other methods of improving reliability, fuzzy extraction techniques and error correcting codes, are more complex and involve sacrificing bits of entropy to produce a shorter but more reliable signature (or inducing predictability with auxiliary helper data). A typical error correcting code is an implementation of the BCH code [14]. Like the uniformity debiasing, while these processing steps can help to enhance the reliability of the PUF at a cost, strong inherent reliability remains a very important consideration when judging the merits of a PUF implementation.

Finally, while not applicable to the individual bit case, once a collection of bits is concatenated into a longer string (forming a full response of the PUF) the error rate splits into the false negative and false positive rate, or FNR and FPR, for the string. This effect is introduced and discussed in section 6.2 when consideration extends to this multiple-bit level.

2.1.2.3. Physical Appraisal

The final field in which PUFs can be evaluated is in the properties of their physical embodiment [15]. The overall physical specifications of a certain complete PUF as fabricated have a very strong dependence on the specifics of that particular implementation (as designed and manufactured) itself, while the statistical evaluation is slightly more independent, and the qualitative properties are much more independent. As such, these properties can vary quite significantly based on factors separate from the particular abstract conceptualisation of PUF itself – for instance, the node size of the evaluation integrated circuit chips or the logic levels of the post-processing communications. Therefore, evaluations of the physical specifications of a PUF should, as well as quoting the overall specifications of a system, attempt to outline and discuss the inherent and theoretical optimal specifications of the PUF design. For instance, take the case of a prototype PUF that employs micro-electromechanical system (MEMS) sensors as a source of entropy, but uses a system of multiple integrated circuits to evaluate. This PUF could be described as having a lower bit density when looking at the PUF system as a whole, rather than the higher bit density entropy source when taken separately from the evaluation bottleneck, and the concept disregarded. Equally, however, one can take an image of a PUF consisting of a nanoscale etched area evaluated using an attached scanning electron microscope (SEM), constructability considerations aside. Here the bit density of the entropy source would be very high, but would be inseparable from an evaluation mechanism that adds a huge additional footprint, and dilutes the bit density of the whole system to being infeasibly small. This could make the concept of using nanometre surface variations appear much more reasonable than it would be in actuality. There must therefore be an amount of nuance and additional consideration when quoting the physical metrics of a PUF. Additionally, a delineation should be made between the physical metrics of a given PUF system as a whole and the physical metrics that the PUF concept can allow for, depending on whether it is the design of a specific PUF or the capabilities of the concept of PUF that is being evaluated.

2.1.2.3.1. Bit Density

The first consideration as to the physical requirements of a PUF is the space required to host the PUF. This is typically normalised by the response bits supported by the PUF and quoted as the bit density, or unit area (or in some cases volume) required per bit – either for the whole system or just the entropy source by itself. This value can also be considered as the entropy density, or entropy per unit area, for instance using the min-entropy accounting for uniformity in the respective section. This is important to ensure that the PUF can be of a reasonable size to integrate into the authentication object, or otherwise manage logistically. Some PUFs, especially intrinsic PUFs such as the SRAM PUF, have a bit density that strongly relates to the progress of the technology to which they are intrinsic. Here the bit density would be proportional to the transistor spacing of the manufactured integrated circuit onto which the PUF is to be based. Other, extrinsic, PUFs may have the evaluation apparatus external and reusable. In this case, the entropy density of the entropy source must be considered by itself, as well as alongside the feasibility considerations attached to the evaluation apparatus, including its own footprint.

2.1.2.3.2. Throughput

The next consideration is towards the timing requirements of the PUF, or how long the PUF takes to return a response. This is typically expressed as the property of throughput, as bits per second – either for the whole system or for the entropy source itself. This is important, as too low a throughput could add a prohibitive delay to the reading process, and in general faster processing is more attractive. For PUFs based on non-electronic sources of entropy this limitation typically exists only with the evaluation mechanism, independent of the entropy source. This is because these PUFs typically examine the reaction of a surface or object with incident optical illumination, a reaction that occurs at the speed of light, with the evaluation mechanism as the only meaningful cause for delay. However, PUFs based on electronic sources of entropy can often find that the entropy source influences the total evaluation time, as the time it takes for the electronic circuit to take a characterisation measurement of the electronic entropy source is slowed by the inherent inductance and capacitance characteristics of the elements under test. Additionally, certain electronic PUFs, such as the arbiter family of PUF concepts, derive their entropy from the timing response of certain circuits meaning a certain latency in signal, varying by higher-level racetrack design, must be anticipated regardless of evaluation time. For PUFs that directly measure the IV response of some electronic component, such as the devices used in this work, the practical throughput limitation would lie with the rate of an evaluative analogue to digital converter component, as opposed to the much higher theoretical limit based on device frequency response. In certain situations, such as with a strong PUF with a relatively smaller CRP space, a lower throughput (or a throughput deliberately limited) may be desirable. This would be to help prevent an attacker from copying a significant portion of the supported CRPs in a reasonable time, aiding the strong PUF's natural resistance to emulation that derives directly from its greater number of possible CRP positions.

2.1.2.3.3. Power Consumption

The final major consideration for the physical metrics of a PUF is in its power consumption, or how much power is needed to run the evaluation. This property is typically expressed as energy requirement of evaluation per bit of measurement, or usable entropy, in joules per bit, but is sometimes separated from the time component to be expressed as power consumption during the evaluation in watts. A larger power requirement not only adds additional cost, but in the case of small-size integrated PUFs can cause issues of temperature regulation if treated without appropriate consideration. This could cause the PUF to unduly heat up, causing damage or at least systemically shifting the responses measured by the evaluation circuitry (should the signature of the entropy have a temperature dependence). Again, like throughput, for non-electronic sources of entropy there is usually not a certain power consumption as an inherent part of the entropy. However, while electronic entropy sources have a certain inductance and capacitance that affects reading time, they also have a certain inherent resistance, and therefore power dissipation, associated with them. If the evaluation apparatus is managed externally and separately from the entropy source the power requirements, like the evaluation footprint, tend to be of less impact.

2.1.3. Types of PUF

Physically unclonable functions can most readily be split up into two overarching domains [16]. The first of these are PUFs that are non-electronically embodied and involve systems such as evaluating the optical, magnetic or RF properties of a source of entropy. An example of this would be evaluating an array of cholesteric liquid crystals by examining the light reflected from their surface with a camera and microscope [17]. The second of the higher-level categories consist of PUFs that are electronically embodied, and as such typically look at the electronic properties of their respective source of entropy, such as resistance [18], capacitance [19], latency [20], and memory cell state [21]. These two groups typically have very different applications, with the former typically acting to authenticate a physical object, such as medication or luxury goods, and the latter to authenticate electronic devices, such as SIM cards or other integrated circuits, or for secure communication and cryptography. As PUFs all operate on the same principle they can, in theory, be used for any application (CRP support notwithstanding), but the different means of evaluation lend themselves to these purposes. Direct electronic connection to the surface of medication packaging is more cumbersome than aiming a camera at the same surface, and attaching the optical entropy source and required camera to every chip-and-pin card manufactured would result in extreme limitation, as an example. Electronic domain PUFs, such as the devices discussed in this work, can be considered as split into three groups, divided by what can be considered the philosophy of the concepts beneath. These groupings are depicted in Figure 1 and described in the following subsections.

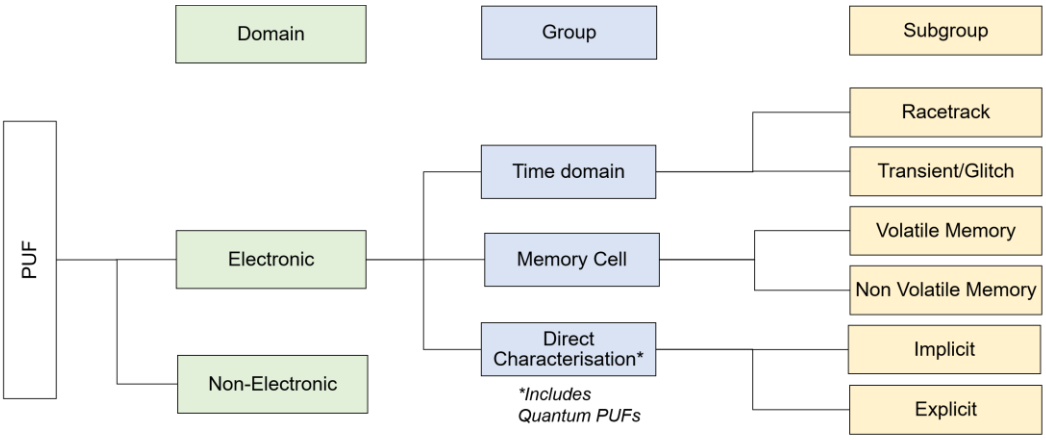

Figure 1: A simple scheme to organise the collection of electronic PUF concepts that currently exist, and to contextualise the RTD PUF of this work. The RTD PUF considered as a part of this work would be classified into the explicit-randomness direct characterisation category, found at the bottom right of this figure. A more advanced organisational scheme can be found in external works [16].

2.1.3.1. Time Domain

The time domain family of PUFs are concepts that rely on the electronic timing, or latency, characteristics of a circuit to derive a unique fingerprint. This group of PUFs can be considered as split into two subgroups. These are racetrack-style PUFs, which look at the variation in latency of defined, repeatable signal traces, and transient or glitch style PUFs, which look at the transient phenomena that emerge from more complex logical systems. The most notable example of racetrack PUFs is the arbiter PUF [22], where two racetrack conductive paths of variable configuration are set as the challenge, and signal first reaching a final sensor is taken as the response. This can be compared to the glitch PUF [23], which takes transient signal spikes that are inherent in the time evolution of a logical circuit as the unique fingerprint for the circuit. Within this time domain grouping, the arbiter and ring oscillator PUFs [24] (examining the alternation frequency of an odd-number ring of inverter gates) see the most industry and academic focus.

2.1.3.2. Memory Cell

The next grouping of PUFs are concepts that employ elements conventionally considered as memory cells as their source of entropy. This group can also be bisected, here into PUFs looking at volatile memory cells and those with a focus on non-volatile memory cells. Volatile memory cells typically look at some measure or reaction of a collection of conventional SRAM [25] or DRAM [26] memory that is, or can be manipulated to be, repeatable and divisible into two response bit states. By examining this property over a whole collection of these common memory types a repeatable fingerprint can be built up that is unique to the collection itself. An example of this would be the SRAM PUF, where each unit of SRAM memory is given power without an applied bias deliberately pushing the ostensibly symmetrical device into either of the bistable states (as is the case in typical operation). Under these conditions, the SRAM cell naturally falls into one state or the other, as a result of fine mismatches between the components constituting either wing. This lends an easy fingerprint, or source of entropy, to the collection of memory cells. Non-volatile memory cells [27, 28], on the other hand, typically operate somewhat differently. These adjust some parameter during the usual memory state assigning process, such as the bias of a memristor [29] or STT-MRAM unit cell [30], or the heating profile of a PCM cell [31], to cause the cell to fall into either memory cell state with equal probability. Unlike responses derived from volatile memory sources, repeating this process to the cell does not give the same outcome each time, and gives rise to a different random distribution as a whole. This random distribution of states is stored semi-permanently in these non-volatile memory cells, and the conventional reading of the memory cell states are taken as the responses for the PUF. The process can be reapplied to create a new response set, making these PUFs typically able to be used in a resettable manner [5]. In this way, this type of PUF can be considered analogous to a form of memory with an inbuilt random number generator acting as key storage and assignment, respectively. In this space the most commonly applied PUF concept is the SRAM PUF, operating as earlier in the section, due to the ubiquity of SRAM itself as a memory source.

2.1.3.3. Direct Characterisation

The final family of all-electronic PUF concepts is those that employ direct characterisation. This categorisation seeks to differentiate these PUFs from those based on the principles of either looking at the latency of a chain of components as with the time domain PUFs or aiming to convert the bistable properties of memory cell systems into fingerprints, as with memory cell PUFs. Instead, these PUFs aim to measure a time-independent characteristic of a certain electronic component or entity of a circuit in itself as a bottom-up approach, in contrast to starting with the higher-level design block of a memory cell array as a top-down approach. This section can be considered as divided into two subsections, this time by PUFs with implicit or explicit randomness sources, as is described in subsection 2.1.2.1.1. In other words, the PUFs are divided into those that use a typical element of the host circuit as the source of entropy, and those that do not. An example of an implicit direct characterisation PUF would be the VIA PUF [32], which manipulates parameters in the typical via interconnect perforating process to produce an array of vias with a 50/50 chance of connectivity, usable as a fingerprint. An example of an explicit direct characterisation PUF would be the coating PUF [3], where a coating of ferromagnetic particles in epoxy is applied to the surface of an integrated circuit. This random distribution of particles results in a unique fingerprint deriving from capacitance variation when measured by a comb of sensors underneath. This PUF has the atypical benefit of voiding the coating entropy source if an attacker attempts to physically access the circuit underneath.

A notable subcategory of direct characterisation PUFs are PUFs that employ quantum phenomena. In this space, there are two approaches – PUFs that use quantum effects to enhance security for a typical, classical (0/1) PUF readout and PUFs that generate a quantum readout themselves. Quantum readout PUFs, such as specially designed optical speckle pattern PUFs, generate a quantum state as a response to a certain challenge, allowing integration with conventional quantum cryptography and mitigating against simulation attacks [33]. PUFs that employ quantum effects for a classical readout exploit quantum effects, typically quantum confinement, to enhance the security of a PUF with conventional 0/1 bit state output. The involvement of the quantum process acts to enhance security, and the rest of the PUF system can be considered in the same way as other, non-quantum, PUFs. An example of this quantum-enhanced security PUF is the PUF in this work [8]. Here the quantum confinement effect of a resonant tunnelling diode is used to magnify and increase the measurement resolution of a collection of atoms, giving a signature on the atomic scale that can be evaluated with macroscopic equipment. This means that copying the source of entropy (and therefore cloning the PUF) requires 3D atomic resolution characterisation and fabrication – a task that is beyond feasibility now and long into the future. Additionally, a quantum-confinement derived negative differential resistance region reduces the risk of simulation attacks, as the feature is hard to reproduce in conventional electronics.

2.2. Resonant Tunnelling Diodes (RTDs)

Now that the background to PUFs in general has been laid out, this section will go on to provide the theory elements for RTDs and their specific use in PUFs.

2.2.1. Semiconductor Background

A resonant tunnelling diode is a semiconductor component that exhibits a voltage-controlled negative differential resistance region [34]. This is a region over which the current passing through the component decreases as the voltage increases, demonstrating an inversed non-linear proportionality compared to most conventional components, which display increasing current transmission with increasing voltage over their functional range. This effect is a consequence of the diode's semiconductor conduction band structure, which can be seen under increasing bias in Figure 2. As can be seen, the device consists of two barriers defining a quantum confined well and a corresponding first quantised energy level within. It is worth noting here that as with all quantised energy levels further levels theoretically exist but here, but as with most resonant tunnel diodes, the next energy level is at a higher energy than the barriers, and so does not have an impact on the RTDs electronic behaviour. An electron incident on the diode can pass through the component in one of two ways: either through tunnelling across the barriers or, should the electron have sufficient energy, 'jumping over' or bypassing the barrier altogether in a process known as thermionic emission. The consequence of the confinement is that only electrons that have an energy corresponding to this energy level can exist within the well. This means that an electron tunnelling through the barrier must either tunnel through the whole structure at once (improbable and negligible), or tunnel through both barriers using the confined energy level as an intermediary. This pathway is more probable but requires the exact incident electron energy corresponding to the quantum confined energy level to proceed.

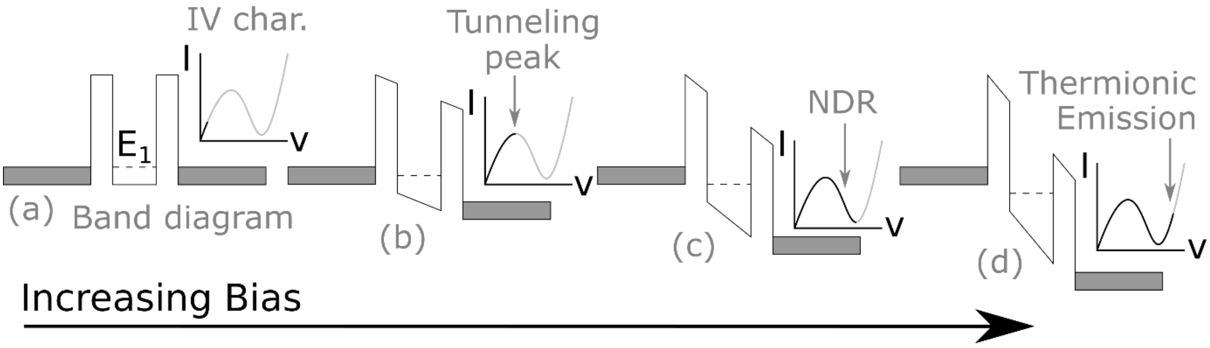

Figure 2: The band structure and consequent electronic characteristic for a resonant tunnelling diode at increasing levels of bias.

The incident electron energy exists as a distribution determinant on the bias applied across the junction, and as such the current-voltage (IV) behaviour of the diode seen in Figure 2 arises through the following steps. As the bias across the junction increases (towards the right on the figure), the average energy of the electrons as they relate to the confined energy level increases. As the average electron energy approaches that of the energy level, an increasing number of electrons randomly have the exact energy required to pass across (Figure 2a). This increase occurs until the average electron energy equals the confined state energy, at which point there is the greatest flow of electrons passing via the tunnelling mechanism, and a peak is formed (Figure 2b). As the applied bias (and therefore average electron energy) increases further, fewer electrons in the distribution randomly have the appropriate energy, and so the current declines and a negative differential resistance region is actualised (Figure 2c). This decrease in current continues until the average energy is such that a substantial and increasing number of electrons have the energy to bypass the double barrier thermionically (Figure 2d). A second, now exponential, positive differential resistance region forms as consequence, resulting in the N shaped IV characteristic fundamental to the RTD. This linking between the small-scale structure/behaviour of the diode and its macroscopically measurable IV characteristics, here arising from the quantum confinement and tunnelling peak, makes the device a very attractive candidate for quantum secure PUFs and RNGs [35].

2.2.2. RTDs as PUFs

There is interest in the employment of RTDs to further the security of a PUF [9]. This is because the exact positioning of the confined energy level determines the macroscopically measurable position of the voltage peak in the RTD IV characteristic, and is in turn highly dependent on the individual configuration of the collection of atoms within the physical region of the quantum well. The quantum confinement can therefore be described as acting to achieve an otherwise unreachable high-resolution measurement of a small group of atoms, and as such while the signature of these atoms is readily measurable, attempts to physically map and recreate the atomic structure required to clone the PUF remain infeasible. The quantum confinement influences the IV characteristic for every area of the tunnelling region. As such a quantum-enhanced signature can be derived from many different aspects of the IV characteristic, but in the form studied in this work the position of the peak is evaluated for the response, due to its direct relationship to the energy level, in the voltage domain based on the infinite well approximation described below, or in the ensuing current at that peak.

$$V_P = \frac{2}{e}\left(E_W - E_C + \frac{h^2}{8m^*L^2}\right) \tag{10}$$

In the above equation, the parameters L represents the width of the quantum well layer, m^* represents the effective mass of the electron in the medium of the quantum well, and where E_W and E_C represent the minimum energy inside and outside the barriers of the diode, respectively. The constants e and h stand for the electronic charge and Planck constant, respectively. V_P denotes the voltage at the tunnelling peak. The voltage value of this peak is preferable to the current value of the peak, as the current domain measurement is much more strongly influenced by factors such as the temperature of the device, which results in a less consistent response [36-38]. Of additional note, multiple RTDs connected in series derive a convoluted IV characteristic, which allows for a signature dependant on the entropy, or quantum well atomic configuration, of multiple diodes. In this way a strong PUF may be derivable, as here a unique signature is dependent on the permutations of a combination configuration of a chain of diodes, and as such has a permutation space that scales with super-linear complexity. This work, however, focuses on the employment of RTDs in a PUF with weak scaling. A discussion on the varying means of extracting signatures from RTDs can be found as part of section 4.

Chapter 3: Experimental methods

As part of this work, resonant tunnelling diodes (RTDs) are electronically characterised and used as the case study for the process of simulating a PUF. These electronic characteristics depend on the parameters of the fabrication process and the conditions of the measurement process, and as such these steps are outlined and detailed in this section.

3.1. Resonant Tunnelling Diode Fabrication:

This subsection aims to elucidate the process by which the RTDs employed in this study are produced. It is worth noting here that the author had no personal involvement with the manufacture or design of these diodes and seeks to discuss this fabrication process for completeness only. A full description of the manufacturing process can be found in external work [39].

3.1.1. MBE

The active region of the RTDs used for this study consist of a vertical stack of different thicknesses of indium gallium arsenide, of varying doping, strain, and composition ratios, along with two layers of aluminium arsenide to define the tunnelling barriers. These layers, defined in Table 1 below, were deposited onto a semi-insulating (100) indium phosphide substrate through the use of a RIBER V100 Microwave Beam Epitaxy (MBE) system.

Layer	Thickness (nm)	Doping Concentration (cm^{-3})	Notes:
n^{+}-In$_{0.53}$Ga$_{0.47}$As	45	2.00 x 10^{19}	Highly doped to allow for ohmic contact
n-In$_{0.53}$Ga$_{0.47}$As	25	3.00 x 10^{18}	Emitter layer
In$_{0.53}$Ga$_{0.47}$As	20	Undoped	Spacer layer (preventing dopant diffusion)
AlAs	1.3	Undoped	Barrier layer (defining quantum well)
In$_{0.8}$Ga$_{0.2}$As	4.5	Undoped	Quantum well (highly strained lattice)
AlAs	1.3	Undoped	Barrier layer (defining quantum well)
In$_{0.53}$Ga$_{0.47}$As	20	Undoped	Spacer layer (preventing dopant diffusion)
n-In$_{0.53}$Ga$_{0.47}$As	25	3.00 x 10^{18}	Emitter layer
n^{+}-In$_{0.53}$Ga$_{0.47}$As	400	1.00 x 10^{19}	Highly doped to allow for ohmic contact
InP	N/A	N/A	Substrate layer

Table 1: The epitaxial structure of the RTDs used in this work

3.1.2. Etching and Metallisation

After growing the active region, contacts were applied to the RTDs and were etched into individual devices. The top contacts consisted of 50 nm of titanium followed by 200 nm of gold and were deposited using thermal evaporation in an area defined by optical lithography. These contacts also served to act as the mask for etching the III/V semiconductor stack into individual devices. This etching was done through a process that first employed reactive-ion etching using a mixture of CH_4 and H_2, and then a wet etching process using an aqueous etchant of phosphoric acid and hydrogen peroxide. Finally, bottom contacts of 50 nm titanium and 500 nm gold were deposited, through the same means as the top contacts. The size of the top contact mask element that defined the cross-sectional area of the device's active region was varied, to produce a collection of devices of active region area varying from 2 to 6 square micrometre in an area of 15 square millimetres.

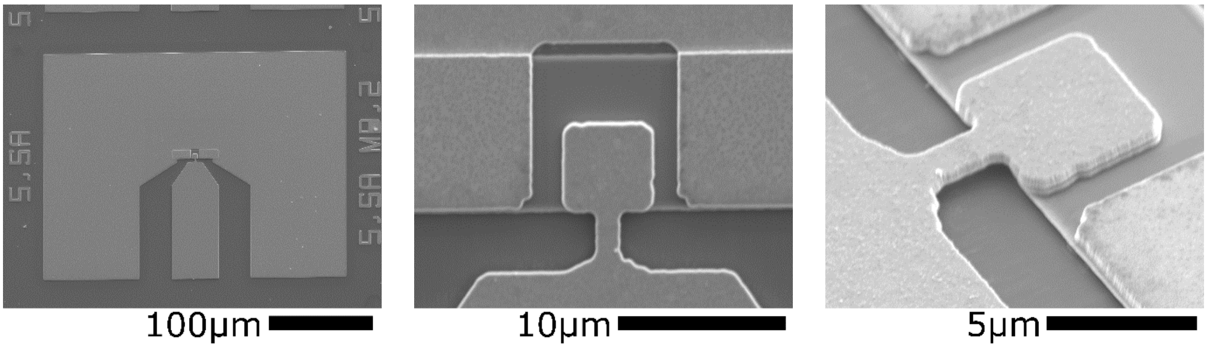

Figure 3: A range of scanning electron microscope images of the resonant tunnelling diode as fabricated, at a range of magnifications. In the leftmost image, the device metallisation is most prominent, whilst the rightmost image focuses on the active region as epitaxially grown by the means outlined in the text.

3.2. Testing Semiconductor Devices

To verify the functionality of the resonant tunnelling diodes before and after processing, a small range of examination techniques were required. These consist primarily of optical examination techniques, to ensure that no signs of visible damage were apparent, and of electrical examination techniques, to ensure the devices worked as anticipated electronically. These electronic techniques were also used to gather valuable data that informed the design and simulation processes described in further sections.

3.2.1. Microscopy Techniques

The RTDs were visually evaluated using a conventional optical microscope (Zeiss Axio Lab. A1) for a high-level view of the device structure, and a Scanning Electron Microscope (SEM) for more detailed imaging. This imaging was performed at various stages of development, for instance to ensure contact surface cleanliness, that the ultrasonic cleaning techniques or use of probes with the probe station were not abrasive to the samples, and to determine the nature of any unexpected electrical behaviour that may result from Electrostatic Discharge (ESD) or overloading, which can result in melting of the active region. The SEM used in this testing was the JEOL JSM-7800F.

3.2.2. Electronic Examination

The RTDs were electrically tested using a SMU (primarily the Keithley 2602B) to source and measure direct voltage and current. This electronic testing apparatus was connected to the sample through the use of a probe station in the initial testing, before moving to wire-bonded chip carriers for less extensive but more reliable and integrable testing. The probe station used in this work was a Wentworth Laboratories SPM197 probe station, employing two micro-positioners mounting tungsten probes of 1.25-inch length and 1-micrometre tip diameter.

3.2.2.1. 2- and 4-point measurements

Conventional electronic measurements of an electronic device or sample can be described as 2-point measurements, that is to say that there are two points of contact between the measurement apparatus and the sample. This measurement setup is where a monitored current flows through a sample via the same 2 wires, across which the voltage drop is examined. While this setup is very beneficial in its simplicity, the measurement technique measures the combination of both the sample in question and the connective wires between the sample and the measuring unit. The in-series resistance in the measurement wires and connections is generally constant and of negligible magnitude; but in certain cases, for instance when using a probe station, an uncontrollable and not insignificant resistance can skew data. In the case of probe station measurements, this would be due to the somewhat variable electrical contact the probes make with the gold contact pad of the RTD devices. As the RTD PUFs intend to derive identity from atomic-scale variation in individual RTDs, ensuring measurement variation arises from the manufacture variation of the resonant tunnelling diodes and not from variation at the time of measurement is paramount. It is therefore very useful to get an idea of the resonant tunnel diode IV characteristics without the influence of measurement variation for comparison, a task which can be completed through the use of 4-point measurement schemes, otherwise known as kelvin sensing.

4-point measurements, as seen in Figure 4, consist of two sets of connections onto the sample, forming two loops sharing only the sample itself. One loop (corresponding to the force connections) sources the voltage difference that drives current through the sample, and measures the current passing along the loop (typically as a current source). The other loop, corresponding to the sense connections, consists of only a voltmeter. As described by Kirchhoff's current law, the sum of the current arriving at a junction must be equal to the sum of the current leaving the junction. Since the voltmeter in the sense loop is taken to have resistance tending to infinite, there is negligible current passing across that loop. This means that all the current passing into the junction from the force loop is directed across the sample. Since there are no other junctions in the circuit, we can say that the current across the sample as part of the force loop is the same value everywhere in the loop, including the ammeter (and the current source). This means that the ammeter reading, or current source value, can represent the current across the sample independently of the resistance elsewhere in the circuit. Additionally, because no current flows through the loop containing the voltmeter, the voltage drop anywhere in the sense loop outside the sample is zero, regardless of the other resistances in the loop. This is because the resistance of an element denotes the proportionality between the voltage drop and the current transmission through the element, and so for any level of resistance a current of zero passing through the element must correspond to a voltage drop of zero. This allows us a measurement of both the voltage drop and the current transmission through a sample independently to any additional measurement resistance, and by varying the (not-necessarily-measured) drive voltage of the force loop, an image of the IV characteristics of the device in isolation can be built. While this technique is elucidative, the use of the additional probes prohibits this technique to be used for the full device measurement study, and so is used primarily in a small study to estimate the effect of this measurement resistance and variation in support of the more comprehensive 2-point measurement data. When this 4-point data was taken, the extra probes and SMU used were of the same specification as with the 2-point measurements.

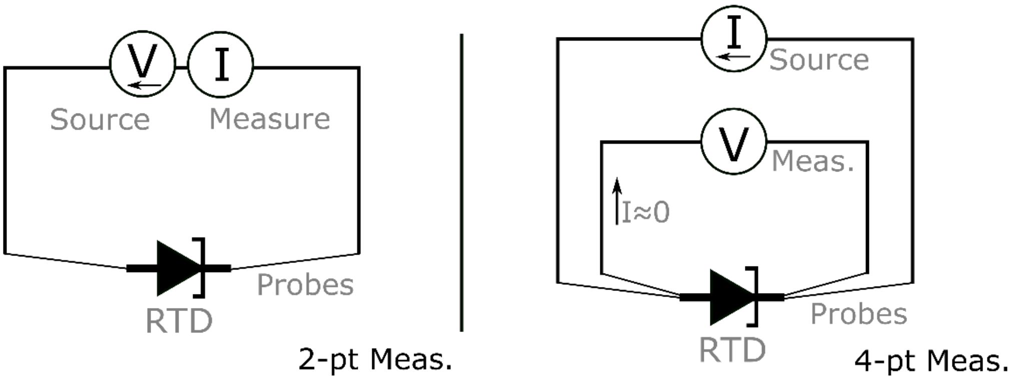

Figure 4: The experimental layout of conventional 2-point and 4-point measurement techniques. The voltage drop due to the resistance of the voltmeter loop probes and wires is negligible, due to the lack of current flow. Since the current is the same at every point along the current source loop, the voltage drops due to wires and probes on that loop do not affect the measurement, are not measured by the voltmeter.

3.2.2.2. Aperture Time

In addition to the configuration of the experimental setup, there were also nuances involved in the operation of the electronic measurement itself that will need tending. The first of these is the SMU measurement integration time of each electronic measurement, often called the aperture time, typically expressed as the number of power line cycles (NPLC) to perform averaging over. When a DC measurement is taken, there is present the effect of noise deriving from the AC power supply line. This noise is periodic and has a periodicity corresponding to the AC frequency of the source power grid – 50Hz, or 20ms period, in the United Kingdom. Averaging over an integer number of cycles is therefore preferable, and so the integration time is described in terms of this value. The fraction or number of these power line cycles, known as the NPLC parameter, represents a trade-off between accuracy and reading time. Additionally, a longer reading time per measurement means a longer period over which the device is being heated, which can also be deleterious to the accuracy of the measurement. This value could change from between 0.01 to 10 power line cycles of integration time, but after testing, it was determined that the optimal value for this measurement parameter trade-off was a single cycle, 20ms integration time or NPLC = 1. This value minimised reading, and therefore heating, time while keeping an integer number of cycles of integration - keeping the measurement independent from the exact time of measurement initiation with respect to the AC power line cycling.

3.2.2.3. Ranging

The next configuration parameter that required management was the ranging of the SMU, for both the voltage sourcing and current measuring elements. An SMU is a device that acts as a source of voltage or current, from a digital input signal to an analogue form (as a Digital to Analogue Converter, or DAC), and conversely from a measured analogue entity to a digital readout (as an Analogue to Digital Converter, or ADC). In both of these cases, conversion between digital and analogue signals occur – a process that can be imagined as consisting of two parts. For a DAC, the first stage converts a digital signal to a value within a certain fixed range, for instance, using an array of resistors bypassed or included based on transistors answering to the digital signal. This value is discrete but variable within a certain range, taking a number of values equal to the resolution of the digitiser (in this example size of the resistive ladder). To improve accuracy, or to access a range greater than the digitiser itself, scaling is then performed. The range (for example 0-10 V in 10 mV steps, at 1,000 discrete points of variation) can be scaled to match a smaller variation at a higher accuracy (for example 0-1 V in 1 mV steps with 10x downscaling) or a wider variation at a lower accuracy (for example 0-40 V in 40 mV steps with 4x upscaling). This scaling factor is known as the source range and represents the trade-off between the accessible range of potential source values and the maximum level of detail by which one can define the source value to be (called the programming resolution, although the smaller range also results in more accurate sourcing). For an ADC the same process occurs only in reverse – the input signal is scaled to a digitiser, a system that for instance combines a DAC resistive ladder with a comparator to determine which two values the input signal stands between. Here the input signal is scaled to best match the ladder's range, using the measurement range parameter, to affect the trade-off between accessible value and accuracy of measurement. This scaling is typically performed with operational amplifiers switching between different feedback resistors, and while by default the SMU automatically switches between ranges to maximise accuracy for any given source or measurement, there is a notable shift (or discontinuity) in measured current when this change happens. This is especially impactful for studies is resonant tunnel diodes, as a downward shift in current can emulate a very small (but very steep) negative differential resistance region, interfering with studies and algorithms that rightly anticipate only one negative differential region per RTD. Additionally, having the same level of accuracy throughout the entire measurement range allows for more reliable uncertainty analysis. The ranging of the SMU must therefore be manually set as a fixed range, typically the smallest range that includes all anticipated source and measure values – an exact value determined by the SMU specifications. For the single-RTD measurements in this work on the Keithley 2602B, the source range was fixed to 0-1 V and the measuring range at 0-10 mA.

3.2.2.4. Compliance

Another consideration when electronically evaluating devices is the compliance value. This is a safety parameter for the SMU sweeping, and (while sourcing voltage) sets the output current value above which the SMU ceases to source voltage, protecting against short-circuiting or device overloading. This is comparable to source ranging insofar as it defines a maximum value for the sourcing of a measurement, but is designed as a safety system rather than a necessity of range/accuracy trade-off. Overcurrent protection is especially important for resonant tunnel diodes due to their exponential thermionic region, since in this region a small increase in voltage can lead to a very high increase in current, and thus a much higher total joule heating experienced by the device. Without appropriate current limiting, therefore, a mistakenly small extra voltage or uncharacteristically early thermionic region can lead to the overloading and destruction of the device. This compliance value varied depending on the RTD size in question (as their current magnitude varies in proportion to size), but in all cases, this was set to slightly higher than the maximum anticipated current value at the tunnelling region peak. For instance, for the $5\ \mu\mathrm{m}^2$devices this was set to 3 mA.

3.2.2.5. Temperature regulation

A final consideration for these measurements was to introduce steps to minimise the effect of temperature on the sample. Any electronic device that is undergoing transmission of current is subject to Joule, or ohmic, heating. While for most purposes this heating effect is negligible or compensated for, in the case of measuring the resonant tunnel diodes attention must be paid towards its minimising. Heating effects are minimised here by allowing the device to cool in-between measurements – that is, performing a pulsed staircase sweep rather than a more conventional monotonic staircase sweep. By minimising the time that the RTD is receiving power (keeping the time allowed for settling and the NPLC at the smallest reasonable value), and by selecting an appropriate time between measurements allocated to powered-off cooling, we can keep measurement events reasonably independent from the measurement order and each other. This value was select to be 1 second during the measurement series featured in this work. The full collection of SMU parameters can be viewed in Table 2 below:

Parameter	**Value**
Aperture time	20 ms (1 NPLC)
Range (Voltage)	0-1 V
Range (Current)	0 – 10 mA
Compliance	3 mA
Sweeping configuration	Pulse staircase sweep (pulse width 1 second)

Table 2: Measurement parameters for the Keithley 2602B source-measure unit used in this work

3.3. Packaging

To integrate the RTDs into electronic devices, they must undergo additional processing. This involves first splitting the full wafer into smaller chips, cleaning these chips, and then attaching them into chip carriers. The RTDs on the chip are then wire bonded to the contact pads of their carrier, and from there can be connected via socket and PCB to a range of electronic testing apparatus and security device prototypes.

3.3.1. Preparation

First, the wafer bearing the full collection of resonant tunnelling diodes was split into smaller chips for easier handling, testing, and integration into electronics. This was performed using a Karl Suss RA120M Automatic Wafer Scriber for the fine scribing, and a manual diamond tip pen tool for some of the more course work. Once divided, these RTD chips were put through a cleaning process. This consisted of a dry-cleaning process through the use of O2 plasma in a Diener Electronics Plasma Asher system followed by ultrasonication, first in acetone and then in isopropyl alcohol, to remove any dust or organic material from the surface before being blow-dried using purified nitrogen gas. The final step in the preparation stage of device processing was to mount the cleaned chips onto chip carriers. This was done by applying a small amount of ‘silverdag’, consisting of an aqueous suspension of very fine graphite and silver particles, drying into a form both lightly adhesive and conductive. The carriers used were Spectrum Semiconductor Materials leaded chip carrier model CCJ02803. This chip carrier consists of 28 contacts, and while maximising RTD capacity by sharing cathodes was considered, the versatility and increased wire bonding yield of having a dedicated pin for the input and output of each mounted RTD was settled upon. Each chip carrier could therefore support up to 14 RTDs, and so the full RTD wafer itself was scribed into chip sizes with that in mind. The probe station electronic measurements discussed prior were performed after the chip cleaning stage and before the wire bonding stage. The mounting of the chips on the chip carriers themselves did not impact the ability to perform these measurements.

3.3.2. Wire Bonding

To connect the diodes into an electronic system, wire bonding must be used. While a probe station can be used to test individual devices, connecting to multiple devices in a way that minimises variation in in-series resistance between devices requires a more permanent bonding solution. This was performed using a TPT HB05 Ultra-sonic Wire Bonder, employing a 25 μm diameter gold wire ball bonded between the contact pads of the resonant tunnel diode and the pins of the chip carrier. The tip used to guide and compress the gold wire was the 19 mm long H61-2/1572-15-750 GM CZ3 capillary.

3.3.2.1. Bonding technique:

To contextualise the parameters that are controlled and optimised for the wire bonding in this work, an introduction to the mechanism of gold ball wire bonding must be made. The method used for electrically connecting the RTD contact pads to the chip carrier was gold wire ball bonding. The wire bonder consists of a spool of fine wire threaded through a capillary tip, with an actuated clamp to feed or retract the wire through the tip. The raising or lowering of the arm that holds the tip and the moving of the stage holding the sample allows for 3-dimensional movement in the placing of the wire. Adjacent to the capillary tip is a small metal arm, called the Electronic Flame Off (EFO) wand, which when activated swings towards the bottom of the capillary tip to induce a spark between the wand and a small amount of exposed wire. This electrostatic discharge causes the portion of the wire that descended from the capillary to form a ball of gold, known as a free air ball. This free air ball is then positioned and compressed onto the surface of the contact pad, and lateral ultrasonic vibrations are applied to the ball through the tip. This movement induces friction bonding, where the gold of the ball rubbing on the contact pad causes localised heating significant enough for the two metals to fuse. Additionally, these vibrations also help to scrub or displace contaminants from the surface of the two metals, to allow for the metal-to-metal coupling to occur cleanly. Once this ball is secured to the first contact, the wire is unclamped and the capillary tip is moved to the site of the second contact point, leading out the gold wire attached to the ball on the way. Once the tip is above the second contact site it descends again, compressing and vibrating the wire pressed by the capillary tip as it crosses from outside the tip as part of the connective loop, to inside the tip and up to the spool. This second bond is known as the stitch bond. The wire is then clamped while being lifted to break free the end of the spooled wire from the now completed bond, and the process can then repeat.

This ball bonding method can be compared to wedge bonding, a process where both ends of the wire are compressed onto the contact with a wedged bond head, with the similar application of heat and ultrasound to bond. This process can be considered as more similar to the second, stitch, bond of the conventional ball bonding process. While wedge bonding allows for a finer pitch than ball bonding, there is no requirement for this in our processes. Equally, the nature of wedge bonding means that a wire must be drawn in a predefined straight line from the initial fixation. Since the wires from the substrate contacts to chip carrier pads are each at different angles and typically not in cartesian alignment with the square sides of the chip carrier, while not strictly prohibitive wedge bonding would introduce many practical challenges. Gold ball bonding was therefore employed for stability, yield, and ease of bonding.

3.3.2.2. Bonding Parameters:

Several controllable variables affect the bonding process. These must be adjusted to achieve bonding and maximise the yield. The first three variables are the capillary downward force, the ultrasonic power, and the time during which these actions are applied. These values are set separately for the first and second bonds. The final variables are the temperature to which the sample is to be heated and the power supplied in the ball generating EFO step, set independently from the bond order. It is worth noting here that some elements of the bonding process were adjusted manually and as such have no defined parameter value. For instance, the height of the bond loop (wire arc between the two contacts) and the length of gold wire descended from the capillary to be subjected to the EFO process (or tail length). Too large a tail length prevents the EFO from creating a ball, while too short a length can either prohibit ball formation or cause the ball to form within the capillary tip, causing a blockage. A tail length of between 400 to 500 µm is recommended, but an exact value for this or the typical loop height cannot be supplied.

The first of these variables, the applied downward force, must be balanced to ensure a good quality bond. If this applied force is too low no bonding process will occur, but a force too large can easily damage the 500 nm gold contact pad of the RTD, causing the whole pad to break apart or lift with the raising bond head tip. The values used for this particular bonding process were found to be around 250 mN for the first bond, and 200 mN for the second, with some variation if the undesirable outcomes of the imbalance described above occur.

The second parameter to be controlled is the ultrasonic power applied to the wire during the welding process. Too low a value and the bonding process will not occur, as with the downward applied force. However, too high a value for ultrasonic power also prohibits successful bonding. This happens by preventing adhesion of the gold ball onto the contact, or reducing the wire subject to the force's ability to deform, causing the wire to snap when being guided away from the first bond. The wire bonder used here employed an ultrasound frequency of 63.3 kHz and was set to use with a relative power value of 250 for the first bond and 40 for the second. Another element to note is a higher ultrasound power usually leads to a wider ball deposited on the contact. This does not cause any issues in terms of bonding pitch due to the arrangement of the contacts on the sample used in these studies, and may in fact help lower in series resistance and increase bond adhesion.

The next parameter that is to be adjusted is the temperature to which the sample is heated. A higher temperature of the sample supports the ultrasonic process, reducing the ultrasonic power requirement for the ball to reach the yield point and fuse to the contact pad. There is no specific downside to increasing temperature, provided the temperature does not melt or interfere with the sample itself. However, since the chip carriers provided very little thermal conductivity between the wire bonder chuck and the sample to bond, the temperature was kept as ambient and not considered as a variable factor for optimisation.

A similar parameter to adjust is the time during which the ultrasound is in effect. This time parameter, like the effect of heating, aids the ultrasound process such that a high activation time (or sample temperature) means less US power is required. This time needs to be kept long enough to allow for the solid-state diffusion inherent in the bonding process, but when set much too high can lead to the same deleterious effects as occur with too high an ultrasound power. The trade-off here is typically between the speed of bonding (not a concern for our non-automated, small-batch purposes) and reliability. A lower activation time can be achieved by a higher US power, and a lower US power can be used if the bonding time is higher. On the whole, a higher bonding time and a lower ultrasound power lead to the most robust outcomes [40, 41]. The bond times used in this study were 50 ms and 200 ms for the first and second bonds respectively.

The final parameter to be managed to ensure a good quality bond is the EFO power. The EFO power relates to the intensity of the EFO spark in forming the free air ball for the initial bond. An EFO power that is too low will produce too small a ball or none at all, while too high an EFO power can cause an overly large, deformed ball or a blocked tip. This value was often varied in response to these symptoms, but a value around 75% of the HB05 model wire bonder's maximum was typical [42]. A list of these parameters can be found in Table 3 below:

Parameter Type	**Parameter**	**Value**
General	Wire specification	25 µm gold
	Tip specification	H61-2/1572-15-750 GM CZ3 (19 mm length) capillary
	Ultrasonic frequency	63.3 kHz
	Temperature	25°C (Ambient)
	EFO power	75%
Initial bond	Bond force	250 mN
	Ultrasonic power	250 (arb. units)
	Ultrasound applied time	50 ms
Follow-up bond	Bond force	200 mN
	Ultrasonic power	40 (arb. units)
	Ultrasound applied time	200 ms

Table 3: Operational parameters for TPT HB05 semiautomatic ultrasonic wire bonder

Chapter 4: RTD measurement

To form a response from a PUF, a source of entropy must be evaluated. This entropy source and corresponding means of evaluation vary between PUFs, but it can be argued that all PUFs start with an analogue measurement of entropy in one way or another (or, more accurately, a typically uncountable large number of some certain quanta) before conversion into bits. However, in some cases, this conversion occurs at a lower, pre-processing, level. For instance, an SRAM PUF cell or arbiter PUF configuration appears to give an immediate binary output, but can equally be considered as having an analogue source of entropy (transistor turn-on characteristic for SRAM, signal latency for arbiter) that is immediately and inherently turned to a single bit state. This is done by comparison to a second equivalent sub-element (the second wing of the bistable SRAM cell, or the competing racetrack signal path respectively), as in the comparative binning technique in section 5. Another ostensibly immediately binary PUF, the VIA PUF, operates on the existence or nonexistence of connectivity within a deliberately impaired via connection. This may appear to be directly of a binary output, but again connectivity (or more accurately conductivity) is an analogue (or typically very finely discretised) scale – a disconnected via still has a very small but analogue and finite conductivity value (or equally a high, existent, and analogue resistance value). To separate this range of conductivity values a threshold, however low, must be chosen (either deliberately in processing or somehow implicitly) to turn this value range into a 0/1 state – or threshold binning as in section 5. This is all to say that even if a PUF is considered as having an immediate binary response at the typical level of examination resolution, it can be broken down into an analogue element with a form of implicit binarisation at the level below. The evaluation of the analogue entropy element for the PUF varies between PUF concepts, as the analogue value taken as the PUF’s signature is different in each conceptualisation. In this section a study will be undertaken to determine the best way to extract an analogue measurement specifically from a collection of resonant tunnel diodes, providing optimal properties for binarisation. Binarisation, however, can take a form generic to all PUFs and is discussed in section 5.

4.1. RTD Data

To allow for a wide collection of PUF instances to be examined, a substantial collection of resonant tunnelling diodes was measured, and signature characteristic values extracted. This distribution of signature values can then be used to procedurally generate the signatures of an arbitrarily large number of RTDs, which can then be grouped into simulated PUF collections, binarized and tested for resultant physical security properties. For a single resonant tunnel diode, the most forthcoming characteristic value is the position in either current or voltage of the tunnelling peak, but also a technique locating another (arbitrary) point on the first positive differential resistance (PDR) region in current or voltage was considered. This peak position can be considered as varying based on uncontrollable variations in the MBE growth process of the device, and to some extent the lithography and other processing stages. In addition, the effect of a range of moving-average smoothing windows was studied, again to find the optimal analogue signature extraction process.

4.1.1. Data Acquisition

The RTD characteristic distribution study consisted of 256 RTDs. These were measured as per the experimental methods section and consisted of RTDs with an active region area of $4\ \mu\mathrm{m}^2$ (136 devices) and $5\ \mu\mathrm{m}^2$ (120 devices). Each of these RTDs was measured at a 101-point resolution in two windows. First, the full RTD NDR and first PDR were measured, in the range 0 to 0.5 V. A second set of measurements, this time in a 50 mV range around the tunnelling peak, were taken - to result in measurements with enhanced resolution around this signature feature as seen in Figure 5. This peak sweep, with its higher absolute resolution, was used as the source of data for the RTD weak PUF developed in this work.

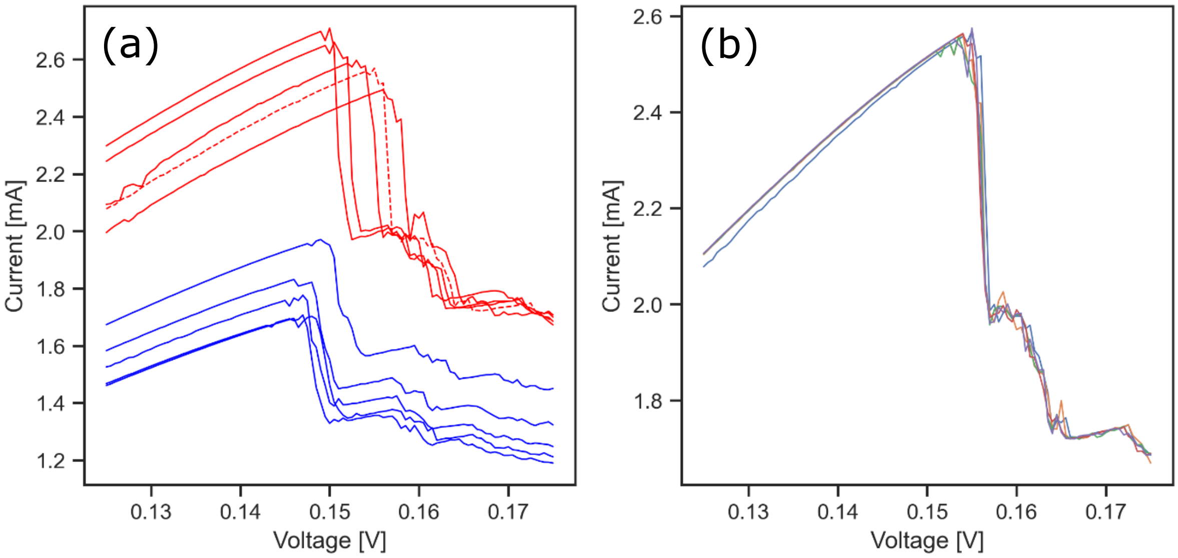

Figure 5: (a) A sample of 10 peak range voltage-current characteristics of the RTD data set. Lines in blue are of 4 μm² size, while lines in red are 5 μm². (b) A sample of 5 peak range voltage-current characteristics measured for the same single RTD (dashed line in (a)).

4.1.2. In Series Resistance

The RTD measurements used in these sections can be considered as the RTD itself in series with a certain amount of ohmic resistance, deriving especially from the measurement probes. This in-series resistance (denoted ISR) may be non-trivial, and the variation of which could add undue variation in peak measurement, deriving from the placing and replacing of the probe station probes. The study into this factor was performed in two stages, first to estimate the average value of the resistance induced by the probes, and then to calculate the variation in that resistance. For the estimation of the resistance value, a series of two- and four-point measurements of a collection of RTDs (using the same probe station) were taken, and their differences compared.

To study the ISR, the two- and four-point measurements of 10 RTDs of $5\ \mu\mathrm{m}^2$ were measured, each a single time with 0.5 mV resolution from around 0.125-0.175 V (101 points). For each voltage point in the lower resistance (4 pt) positive differential resistance region (initial incline, as highlighted in Figure 6a, and as such excluding the less stable peak-most points) the current value was taken. Next, through interpolation, the voltage value corresponding to that current was found for the rightmost (2 pt) IV spectrum. This allowed for a dataset consisting of voltage transformation caused by resistance for around 60 current point values for each RTD. By dividing this difference in voltage by the current, the difference in resistance is uncovered. Expressed most simply this equation would be:

$$\Delta R = \frac{\Delta V}{I} \tag{11}$$

By performing this technique over the positive differential resistance regions of the 10 RTDs and averaging, an estimation of the ISR was be found be to $1.21 \pm 0.04\ \Omega$. While small compared to most deliberate introductions of resistance through resisting elements, this is around 100 times more resistance than can be expected with, for instance, a banana jack style connector. This leads to a shifting of the peak voltage location by about 3.6 mV or around 0.25% of the voltage value itself. The shifting effect due to this extra ISR is not enough to cause a noticeable negative influence on the operation of the design but is beneficial to know and to verify, especially in combination with the variation of this resistance – which, if non-trivial, stands to have a lot more influence.

At an average peak current of 2.60mA for $5\ \mu\mathrm{m}^2$ devices (determined in section 4.2.1) the $3.53 \times 10^{-2}\ \Omega$ resistance standard deviation in ISR (quoted as $0.04\ \Omega$ in prior context) would translate to a $\sigma_{ISR} = 9.18 \times 10^{-5}$ V standard deviation in ISR voltage, using Equation (11). This would act to erroneously increase the inter-RTD voltage variation that is measured since this voltage variation would be the combination of this ISR variation and that of the RTD itself. An estimation of as-measured inter-RTD standard deviation is $\sigma_T = 2.62 \times 10^{-3}$ V (using the 'maximum' extraction method in section 4.2.1 and before normalisation). Subtracting the variance of the ISR from the inter-RTD variance as measured (using the relationship $\sigma_T^2 = \sigma_{ISR}^2 + \sigma_{RTD}^2$) would result in a standard deviation arising from the RTD alone as $\sigma_{RTD} = \sqrt{(2.62 \times 10^{-3})^2 - (9.18 \times 10^{-5})^2} = 2.62 \times 10^{-3}$ V to 3 significant figures. We can therefore conclude that the effect of in-series resistance variation on the measured variation of the RTDs themselves is negligible. It is worth noting that this variation is derived from only 10 RTD devices, and so may not provide the most assured value due to the limited data set. Finally, the average variation in measured resistance across the approximately 60 current slices for each RTD point was around 4.2×10^{-3} Ω, which helps confirm the 2-point to 4-point IV difference can be ascribed to a consistent linear ohmic relationship between the states. In the following work, the absolute in-series resistance is compensated against. This is done by calculating and subtracting the voltage offset caused by the ISR at any given current value, the effects of which can be seen in Figure 6b.

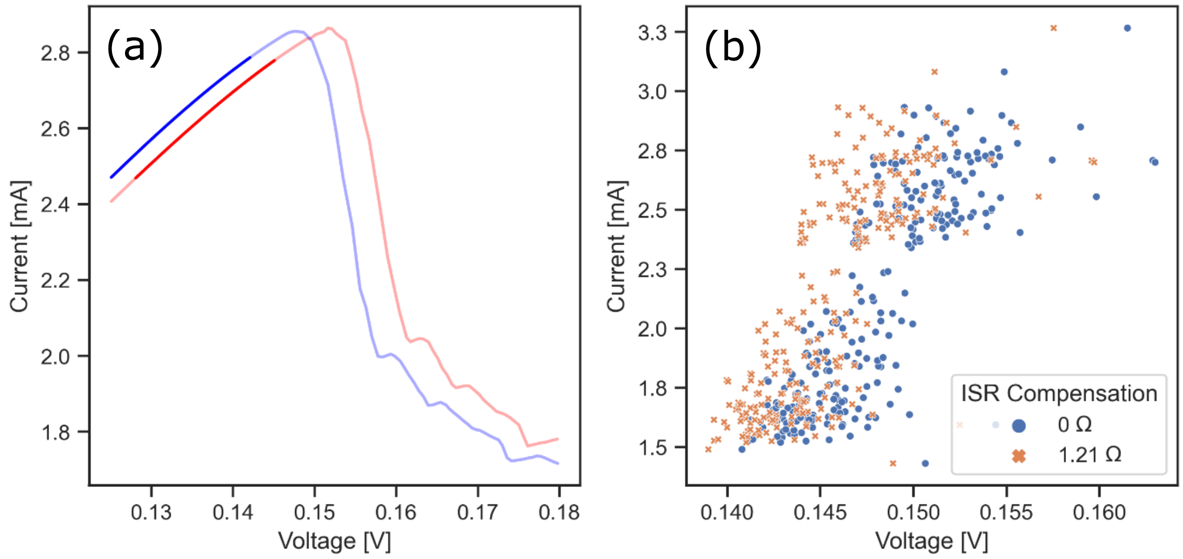

Figure 6: (a) A comparison between a single RTD at 2-pt (red) and 4-pt (blue) measurement modes. Highlighted regions were isolated and contrasted to evaluate in series resistance. (b) The unnormalized distribution of RTD current-maximum values with (red) and without (blue) ISR compensation.

4.2. Evaluation Techniques

Now that the IV characteristics of the RTD dataset can be adjusted to account for the resistance from the measurement apparatus, it is necessary to determine the optimal means to derive an analogue signature from the measurements. In other words, the aim is to reduce the multi-valued IV sweep data into a single value for further processing. This section will look at a collection of techniques for extracting this signature specifically for resonant tunnelling diodes. Evaluations based on more subtle features of the characteristic, for instance the gradient (resistance) at a certain point, the full-width-half-maximum of the peak, or the induced amplification (plateau) level, may have more optimal quality metrics, but would be less directly related to the quantum confinement and typically be harder to evaluate. For these reasons, more speculative features to focus on will not be included. There may also exist dynamic feature signatures or evaluations, for instance latency or LC properties, that would not be apparent from the (ideally) time independent IV sweeps taken here. Finally, effectiveness depends on RTD and measurement specifics, and so this evaluation would only be for the 256 RTDs used in this study, measured at the resolution and range as per section 5.

To compare these techniques, a performance metric must be developed. For this, we use the ratio of the inter-measurement (between RTDs) standard deviation over the intra-measurement (measurements of the same RTD) standard deviation for the mean-normalised values. In this work these entities are given the variable names σ_{intra} or σ_i for the intra-measurement standard deviation and σ_{inter} or σ_o (as in 'outer') for the inter-measurement standard deviation. The ratio between these is given the name σ_{ratio} or σ_r, and sometimes a superscript V or I is included to make clear which domain (voltage or current) is being evaluated. This ratio appears repeatedly in the binarisation formula in section 5, and (assuming measurement resolution is not an issue) quite apparently represents the extraction techniques performance in producing distinct readings for each RTD. The higher this ratio the more separated the true-value measurements are for each device, and also the more concentrated the repeat-measurement clusters are, resulting in lower error rates as fewer measurements stray enough to be mischaracterised.

4.2.1. Maximum

The first and most straightforward way to extract the position of the tunnelling peak is to directly take the data point of maximum current value, in the region in which the peak is likely to exist, as the true location of the peak. This is a simple and direct method to determine the definitional location of the peak (or maxima) in the tunnelling region as measured. However, it is diminished in performance metric due to the presence of noise. The NDR of the RTD acts as an amplifier for incident nonconstant signal, including electronic noise, resulting in an amplification of noise translating to both the apparent current plateau and measured spikes of current in that region. As this region begins just after the true peak of the diode, a lower post-peak current value can combine with the noise spiking behaviour to easily and sporadically be measured as of higher amplitude than the real peak or anywhere on the first PDR. This means that at higher resolutions the maximum current value as measured can be found at varying positions on or at higher voltage and higher currents to the real peak, broadening and imposing a limit on the value for the intra-measurement standard deviation for any given value extraction. This issue does not affect the variations in inter-measurement distribution, and can easily be seen about the peaks in Figure 5 earlier in this chapter.

Applying this technique to the RTD dataset finds the mean value, found as the factor for later normalisation, of the RTDs as 0.143 V, 1.76 mA for $4\ \mu\mathrm{m}^2$ devices and 0.148 V, 2.60 mA for $5\ \mu\mathrm{m}^2$ devices. The observations from this first set of measurements would be that the two sizes of devices should theoretically have the same average position of voltage peak. This voltage difference could result from manufacture variation, a higher than estimated (or inconsistent) in series resistance or the steeper post-peak NDR allowing a greater average voltage shift from the noise issue above. As both data subsets are to be normalised by their respective current and voltage these differences would have negligible impact, regardless of instigation. The more important figures of merit, the aforementioned measurement variations, are found to be $\sigma_{inter}^{V} = 1.77 \times 10^{-2}$; $\sigma_{intra}^{V} = 4.58 \times 10^{-3}$, resulting in a ratio of $\sigma_{ratio}^{V} = 3.88$ for voltage and be $\sigma_{inter}^{I} = 8.83 \times 10^{-2}$; $\sigma_{intra}^{I} = 3.07 \times 10^{-3}$, resulting in a ratio of $\sigma_{ratio}^{I} = 28.8$ for current. In other words, for any confidence interval, the possible values in the measurement of a single RTD are just over one quarter and just under one-thirtieth of the range for possible values for all RTDs in the measurement set for voltage and current respectively. The distributions and kernel density estimation of these peak values, normalised and zeroed around mean values by device size, can be seen in Figure 7.

A final element to consider when evaluating these techniques is the measurement cost of operation. This is to say how many data points need to be measured (or remeasured) to acquire the analogue value (or repeat measurement) taken forward. While this depends on the specifics of the evaluation implementation when practically employed, it would be beneficial to come up with an estimate of this value so that the techniques here can be more adequately compared. A later technique that takes only the current value at a set voltage, for instance, may require only 1 measurement (DAC and ADC cycle) to perform, while another may require the full 101 points taken here to fit an arbitrary function. Even if this latter technique performs better in terms of inter/intra-measurement metric, it will take more DAC/ADC cycles and therefore time compared to the former technique. This time spared due to the faster measurement process is a merit of itself, but can also be reapplied to repeat measurements to enhance the final performance of a PUF employing that technique. Therefore, a certain performance standard may be more effectively achieved with a technique that performs worse in a single instance, should the technique have a much smaller measurement cost. This consideration is fully explored in section 6, when systemic (repeat measurement) error correction techniques are introduced and evaluated, but for now, only the estimation of measurement costs will be considered. In this work, it will be assumed that the measurement resolution employed in a final PUF is the same as the measurement resolution taken for this data. It will then assume that the first measurement of a single RTD requires the full 101 measurement points to determine the position of the peak in question over the full measurement space, but subsequent measurements can be confined to an area around that peak that depends on the specifics of the technique in question. For this maximum method, this region will be taken as 3σ around the peak position, which when calculated from the intra-measurement standard deviation found earlier finds a repeat measurement region of 28 points.

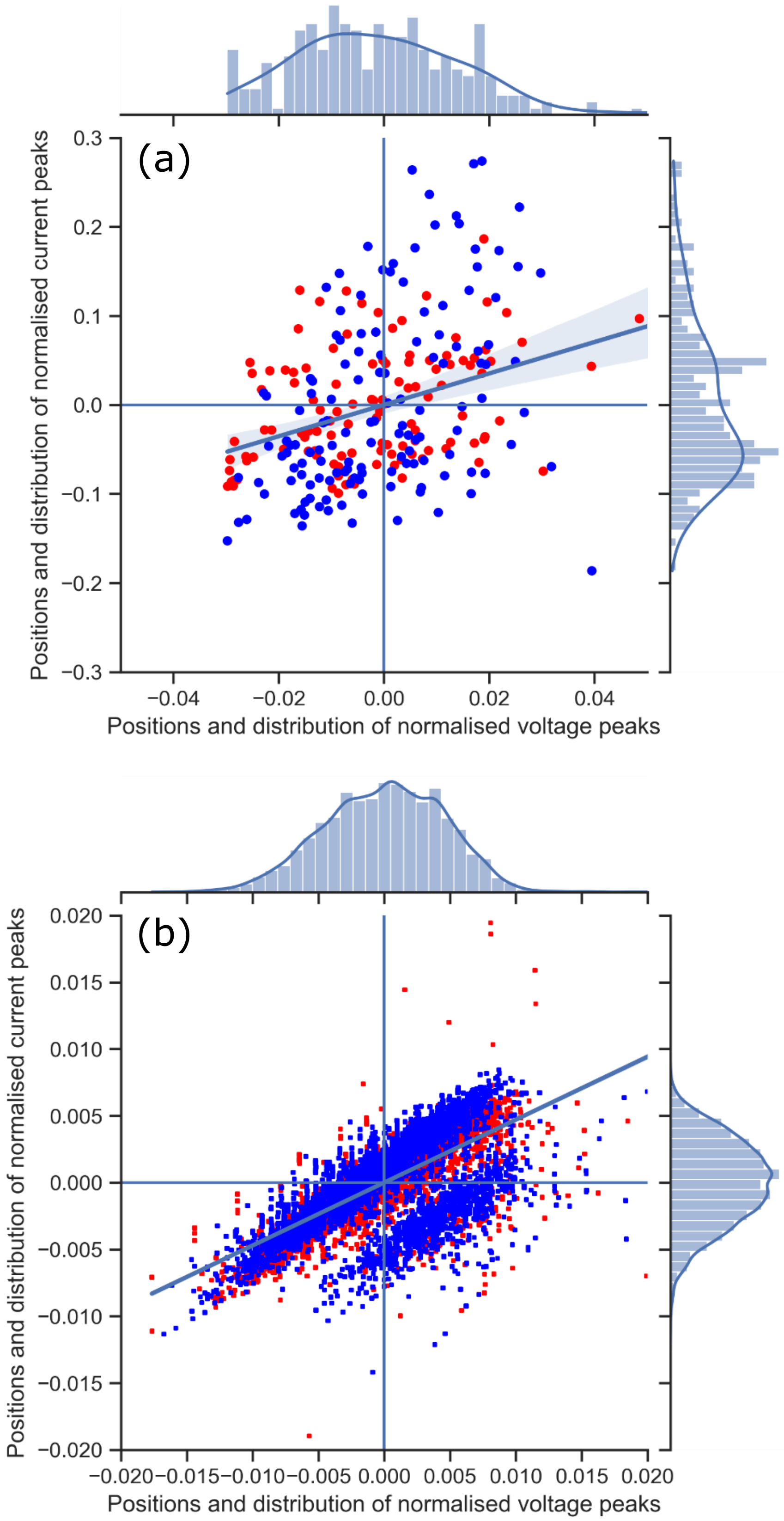

Figure 7: (a) The normalised distribution of (50 points averaged) measurements for each different RTD in the data set using the maximum method (inter-measurement distribution). Included is the linear regression plot for the two variables, as the blue diagonal line and 95% confidence interval shading. (b) The normalised distribution of all measurements around the mean of each different RTD in the data set using the maximum method (intra-measurement distribution). In both plots, the red dots represent the $5\ \mu m^2$ devices and the blue dots represent the $4\ \mu m^2$ devices. Linear regression also included.

4.2.2. Declination

The first adjustment in the attempt of mitigating this misleading noise-spiking behaviour (and improving performance metric), would be to look at the first point in the decline instead of the raw maximum. As the noise-spiking occurs at some point after the consistent true peak, there is often an initial decline before an absolute-greater spike in current is measured. This can be seen in Figure 5. In this way, the decline-locating method means to instead find the first turning point in current, rather than basing it on a raw largest current value which may come later. This method is of comparable complexity to the maximum-point method, here iteratively comparing each data point to the value(s) directly beforehand, rather than to a maximum value.

The first and most obvious implementation of this would be to locate the point preceding the first single declining current value from left to right (therefore starting first at the more stable PDR region). This single point decline can be described as having a decline parameter of 1. When applied, it can be seen in Figure 8 that while this method helps dampen the effect of the noise-spiking, it introduces more intra-measurement variation than it would otherwise reduce from correcting this issue. As can also be seen in Figure 5, while noise in the NDR is amplified, measurement noise, or variation, still exists in the PDR. This means that, occasionally and at higher resolutions, a point in the PDR can be found as erroneously lower than the one before. This can lead to the 'peak' value being found prematurely at various points along the PDR, to the extent that by looking at the intra-measurement distribution one can easily see the trace of the general IV line shape in premature peak readings. While this happens at a lower rate than peak noise, there are more PDR points for this issue to occur over and the occurrences are naturally much further from the peak. This means that while the inter-measurement deviation remains roughly the same as with the maxima method (and the means only slightly downshifted), the intra-measurement variation is greatly increased, resulting in a poor performance ratio of $\sigma_r^V = 0.968$ and $\sigma_r^I = 5.63$. To contextualise this, this means that for any given confidence interval, the voltage measurement values of a single RTD exist in a greater range than the set of (50pt average determined) RTDs themselves. Functioning with similar measurement requirements, the declination method can be considered to have a repeat measurement cost comparable to the maximum method, of 28 points. While very undesirable, this method is still technically viable by repeat measurements as in section 6. This schema can be methodologically improved in two ways – either by manually constraining the evaluation window around the estimated peak (as was done with the 'hybrid' section below) or by increasing the number of consecutive points required to confirm the peak location as follows. Another strategy is applying moving window averaging to the IV data, but this applies to all methods and is instead discussed in subsection 4.2.6.

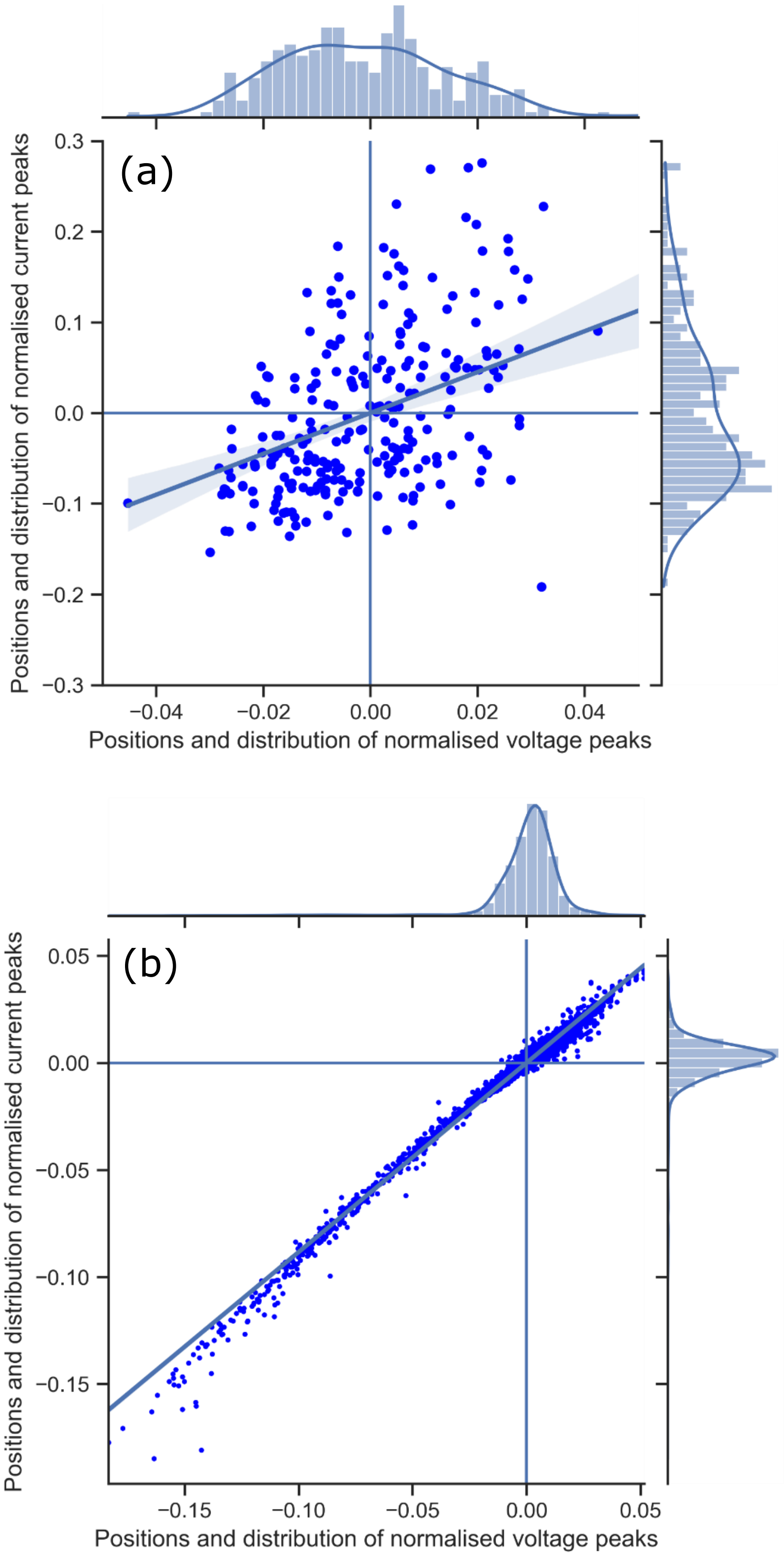

Figure 8: (a) The inter-measurement distribution and linear regression for the single point decline method, for both $4\ \mu m^2$ and $5\ \mu m^2$ RTD devices. (b) The intra-measurement distribution and linear regression for the single point decline method, for both $4\ \mu m^2$ and $5\ \mu m^2$ RTD devices.

The next step in applying this method would be to try taking the location of the tunnelling peak as the point preceding two consecutive declining points (a decline parameter of 2). This circumstance is less likely to be seen in the measurements of the PDR and removes the premature peak-finding issue with no significant increase in complexity. This results in a much lower intra-measurement variation for again a comparable inter-measurement variation, leading to a better performance metric than the decline parameter 1 implementation. However, this implementation still runs into problems where, for instance, a noise peak within 1 mV (or 2 points) of the true peak causes an upwards turn in the current (even if not a maximum) that interrupts the expected consecutive chain. This shifts the measurement erroneously to a higher voltage and can be seen by the abundance of points showing the pattern of the NDR past the true peak in the intra-measurement distribution in Figure 9. Overall, the benefit of finding the correct peak when a noise spike maximum is 2 points beyond the true peak is counteracted by the shifting that occurs when any upward turn is within those two points, leading to a performance metric of $\sigma_r^V = 2.66$ and $\sigma_r^I = 7.94$. This is significantly better than the performance of the single-decline method (by just under 3 times in voltage) but still does not exceed the performance of the basic maximum-finding metric. As would be expected, increasing the decline parameter beyond 2 leads to reduced ratio metrics as the right-shift happens further, at $\sigma_r^V = 2.40$; $\sigma_r^I = 3.30$ when requiring 3 consecutive points, and $\sigma_r^V = 1.04$; $\sigma_r^I = 1.61$ for 4.

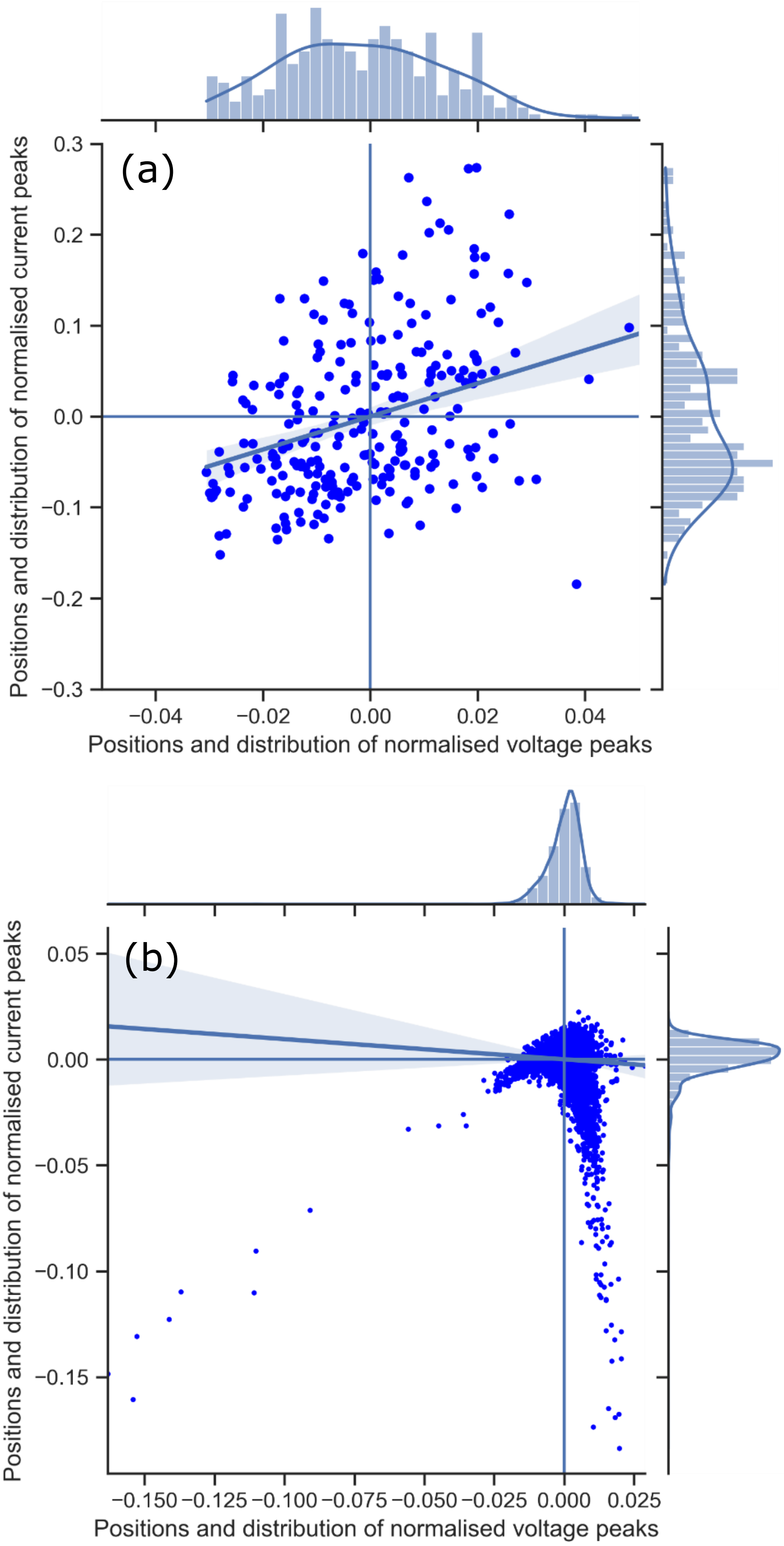

Figure 9: (a) The inter-measurement distribution and linear regression for the 2-point decline method. (b) The intra-measurement distribution and linear regression for the 2-point decline method.

4.2.3. Hybrid

In the pursuit of an analogue value extraction method that exceeds simply taking the maximum point in the performance ratio, the declining method and maximum method were here combined. In this hybrid method the point preceding decline was taken but now confined to a window centred on the maximum (although, given the guarantee of a decline after the maximum, this equivalates to a window only before or on the maximum the single consecutive decline method). This allows the declining method to operate only in the region that would contain the peak and as such limit the premature erroneous locating of the peak along the first PDR. To employ this, the optimal window size must be determined. First, for the single decline method, the optimal meaningful window size was found to be 2 mV (5 pts), or ± 1 mV (2 pts) centred around maximum for voltage and 3 mV (7 pts) for current. A window lower than this is the same as the maximum (as the point approaching the maximum must always be an incline, and the point after always a decline), and higher than this was found to incrementally add intra-measurement variation due to the variety of possible points, rather than ever reducing due to fidelity of measured points. The performance of this combined method with decline parameter 1 was $\sigma_r^V = 3.51$; $\sigma_r^I = 21.0$, which is an enhancement on all implementations of decline seeking by itself, especially in the current, but again did not supersede taking the maximum directly. Applying this technique to two consecutive declines finds an optimal window of 7 mV (15 pts), or ± 3.5 mV (7 pts) about the maximum for current and voltage distribution. This window is high because in some noisy distributions two consecutive declining points do not occur until a significant distance from the peak, and so is needed to include all (or here at least 99.7%) of the RTD measurements. A smaller window size would implicitly exclude the most deviating values, and so misleadingly increase performance as a consequence. A study could be performed into the trade-off between window size, value exclusion and performance metric, but over 35.9% of values would need to be excluded (as a window of size 9, or 4 mV) to gain a performance metric ($\sigma_r^V = 3.48$; $\sigma_r^I = 25.2$) better than the single decline method. At this higher window size, this extraction method has a performance metric of $\sigma_r^V = 3.14$; $\sigma_r^I = 8.95$, which is higher than its nonbounded equivalent but not competitive with other implementations so far studied. Again, this hybrid method has a repeat measurement cost equal to the maximum and declination method, of 28 points.

4.2.4. Projection

An alternative methodology to derive a signature from the RTD tunnelling region would be to forego the pursuit of locating the peak, and instead look simply at the current at a certain voltage in the tunnelling region and vice versa. This method, relying on linear interpolation between the data points, would bypass issues relating to noise measurement around the peak and is here called the projection method. In choosing the values for the projections against which the IV function will be made, certain factors must be considered. First, to ensure a reasonable value, voltage and current projections must be made within the first PDR, in the tunnelling region (in other words values lower than the peak). If the voltage applied is above the peak value it resides in the negative differential resistance region, and so the current measurement is subject to the high variation arising from amplified noise. More profoundly, however, an applied current that is higher than the peak would have no corresponding voltage measurement value in the tunnelling region, and would instead be projected onto the exponential region curve far at a much higher voltage. As the peak positions for the RTDs varies, projection values must be chosen such that the current or voltage is below the peak in all (or at least the vast majority) of cases, providing an upper limit for what these projection values can be. On the other hand, it is the case that all RTD line shapes start from the origin at zero voltage and current and head toward the tunnelling peak, where the greatest inter-measurement variation based on the RTD atomic-scale differences is exhibited. This means that the closer to the peak that a measurement is taken, the more meaningfully different each RTD can be measured to be (absolute intra-measurement variation also increases with increasing measurement value, but only proportionally as expected, and so not an issue). It therefore stands to reason that the optimal projection levels to set are the values closest to (but smaller than) the peak that accommodate the full range of RTDs.

As an aside, this relationship between measurement position and uncertainty raises the uncomfortable point that the variation in values for the measurements or positioning of any RTD using this method is a direct function of that RTD's peak position itself. That is to say that a peak at a higher current would find the catch-all projection value closer to its base, where the unique elements of the RTD have less effect on the measurement. This leads to uncertainty estimates that vary across the distribution of devices, and cannot readily be treated as constant. Basing the projection values on each RTD individually to mitigate this would need to be based proportionally on the measurement of the RTD's peak through another method, and as such would never show better performance than that of the peak-finding method itself.

To most easily get an estimation of the optimal projection voltage and current levels, one can look at the peak distribution information displayed in Figure 7. This uses the maximum-value method and is the most accurate representation of the true peaks that have been evaluated in the series of methods as part of this work. From this, we can see that there is a negligible proportion of RTD peaks found (scatter points) or expected (kernel density estimate) below factor 0.2 (20%) of the mean for current or factor 0.04 (4%) for voltage. We can apply these factors to the means also found in the 'maximum' method section above, to determine an effective configuration of projection values for current and voltage at both device sizes (rounded down), as can be seen in Table 4.

	Size 4 μm^2	Size 5 μm^2
Voltage (V)	0.137	0.142
Current (mA)	1.40	2.08

Table 4: Voltage and current projection levels for the 4 and 5 square micrometre RTDs sampled in this work

This method gives a performance ratio of $\sigma_r^V = 2.83$ for the voltage value at the above-defined current, and of $\sigma_r^I = 35.4$ for the current value at the aforementioned voltage. It is worth noting here that due to the higher variation in current, and the limitations of the data set used, only 216 out of 256 of the RTDs were included when measuring the voltage as a function of current projection (the full set could be included for vice versa). The RTDs excluded had a minimum current that was above the current projection value, and so could not return an accurate measure for voltage. This would result in an artificially inflated ratio value as outliers would be excluded, as with the more stringent hybrid method criteria. Similarly, while the voltage position of peak tunnelling does not change with temperature, the voltage position where a certain current value is measured would do - in a reduced but more complex way. This temperature dependence removes one of the most significant advantages of following the typically lower current value. Overall, however, comparing the performance found here to the maximum method for voltage finds the latter still more effective, rendering this implementation non-optimal and these points moot. One final consideration that affects both projection styles is that this method requires all, or a significant number, of the RTDs to be measured to determine the optimal projection levels. This may add a certain measurement cost of enrolment into this method, as with threshold binarisation in section 5, which would depend on the consistency of the mean and variation at the point of manufacture. After this, however, the measurement requirements for evaluation drop far below the peak-finding methods, requiring only one measurement thereafter and so this technique offers value from this alone.

4.2.5. Gaussian Fitting

A more advanced method for determining peak position would be to fit multiple data points to a well-chosen arbitrary function and deriving the peak values from this rather than the data. This style of extraction method is much more computationally intensive than the methods so far detailed, and doesn’t necessarily seek to exactly locate the peak as per the data, but has the benefit of taking information from more points than just the peak or a single projection value – minimising the effect of deviation in any one point. The most apparent method in this effort would be in fitting a Gaussian, or normal, function to the data around the peak. Despite not defining the exact same shape as the data, this function was chosen for its simplicity and ubiquity. Even more importantly, this fit style was chosen due to the fact that the more parametric degrees of freedom a function can have, the more metastable points a fitting algorithm would observe. This would lead to a higher chance of multiple different interpretations into what exact curve parameters best fit the data, leading to measurements of the same RTD risking increased (intra-measurement) variation. Skipping ahead somewhat, an example of this phenomenon can be seen in Figure 11, where an ‘island’ of peaks exists at a higher voltage and lower current. This can be ascribed to a second metastable interpretation of the least-squares parameter fitting solution. The gauss function is defined as:

$$I(V) = f(x) = H + a \cdot e^{\left(-\frac{(x-b)^2}{2c^2}\right)} \tag{12}$$

Where e represents the exponential function, x represents the voltage and the other variables act as fitting parameters. From this we can see that the peak voltage is defined as simply parameter b, while the peak current (where $x - b = 0$) is the sum of parameters H and a. To standardise the starting locations for the least-squares fitting algorithm, x and c were taken as the voltage-weighted means and standard deviation in current (as relating the function to the normal PDF form), and where H and a were taken as the minimum and maximum current values of the data set, respectively.

With the starting fitting parameter conditions based on the data itself, as above, the only further consideration into operating parameters is the window across which this fitting function should be applied. In the data taken here, and for any non-adapting truncation of voltage range, the number of PDR points, as compared to NDR points, would vary dramatically due to differences in centring, shifting the curve to overly accommodate, for instance, the inclining points should they exist in higher proportion than the declining. Similarly, as can be seen in the sample data in Figure 5, a plateau forms as an effect of noise amplification at a small distance beyond the peak itself. Algorithmically fitting to these horizontal points would shift the final gauss fit, with a mean and induced variance again dependant on how much of this plateau is included in the fitting window. For these reasons, the RTD-measurement range passed to the fitting function is to be of a carefully managed voltage span, and roughly centred on the peak through the use of the maximum method as before. To determine the optimal window size for the fitting of the data, the extraction process was applied to a range of window sizes, and the performance ratio taken for each. The plateau effect occurs well within 15 mV (30 pt) of the peak, giving an upper bound of fitting range span, and one can assume that more than 5 data points are required for an adequate fit, giving a 2 mV lower bound. The results of this study are shown in Figure 10 below.

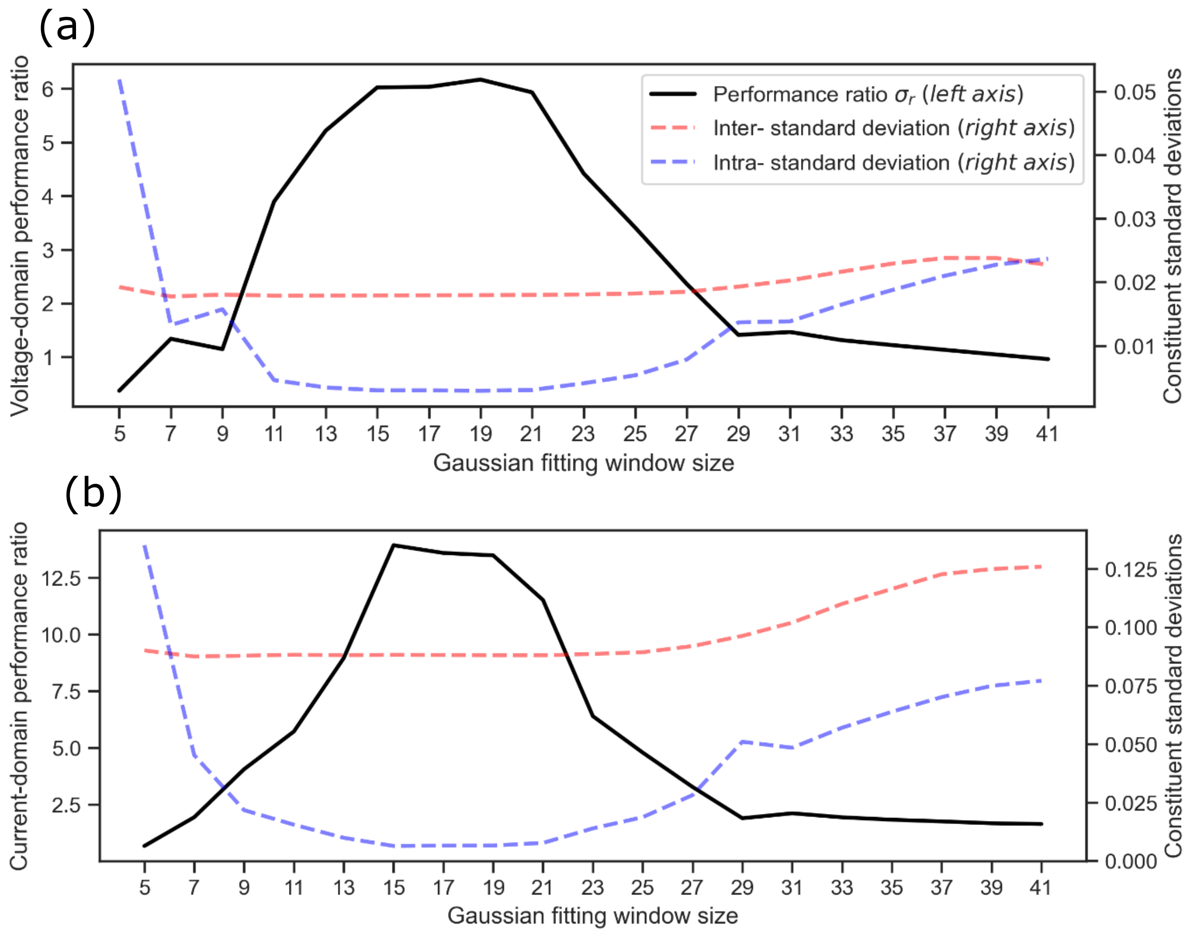

Figure 10: (a) The relationship between fitting window size and configuration performance for the Gaussian fitting method in the voltage domain. (b) The relationship between fitting window size and configuration performance for the Gaussian fitting method in the current domain. In both figures, the black line represents the performance ratio based on the left y-axis scale, while the blue and red dashed lines represent the constituent intra- and inter- standard deviations, respectively, on the right y-axis scale.

From this figure, we can see that the optimal window size for these devices and measurement specifications is 9 mV (19 points) centred around the maximum value for voltage and 7 mV (15 points) for current. However, due to the small difference between the ratios in the two voltage points, and the value of window-size-derived repeat measurement instances, both will have the optimal trade-off with a window size of 7 mV (15 points). These ranges make optimal the trade-off between data size and inclusion of data detrimental to accurate fitting. Overall, the Gaussian fitting method achieves performance ratios of $\sigma_r^V = 6.17$ and $\sigma_r^I = 13.9$, and can be seen in Figure 10. These techniques can be considered as requiring the number of repeat measurements of the typical 3σ range of peaks (28, as earlier), plus an extra 14 points to account for the expanded window of 7 points beyond and before the peak required to account for the fitting process (for a total of 42 DAC/ADC cycles of repeat-measurement cost).

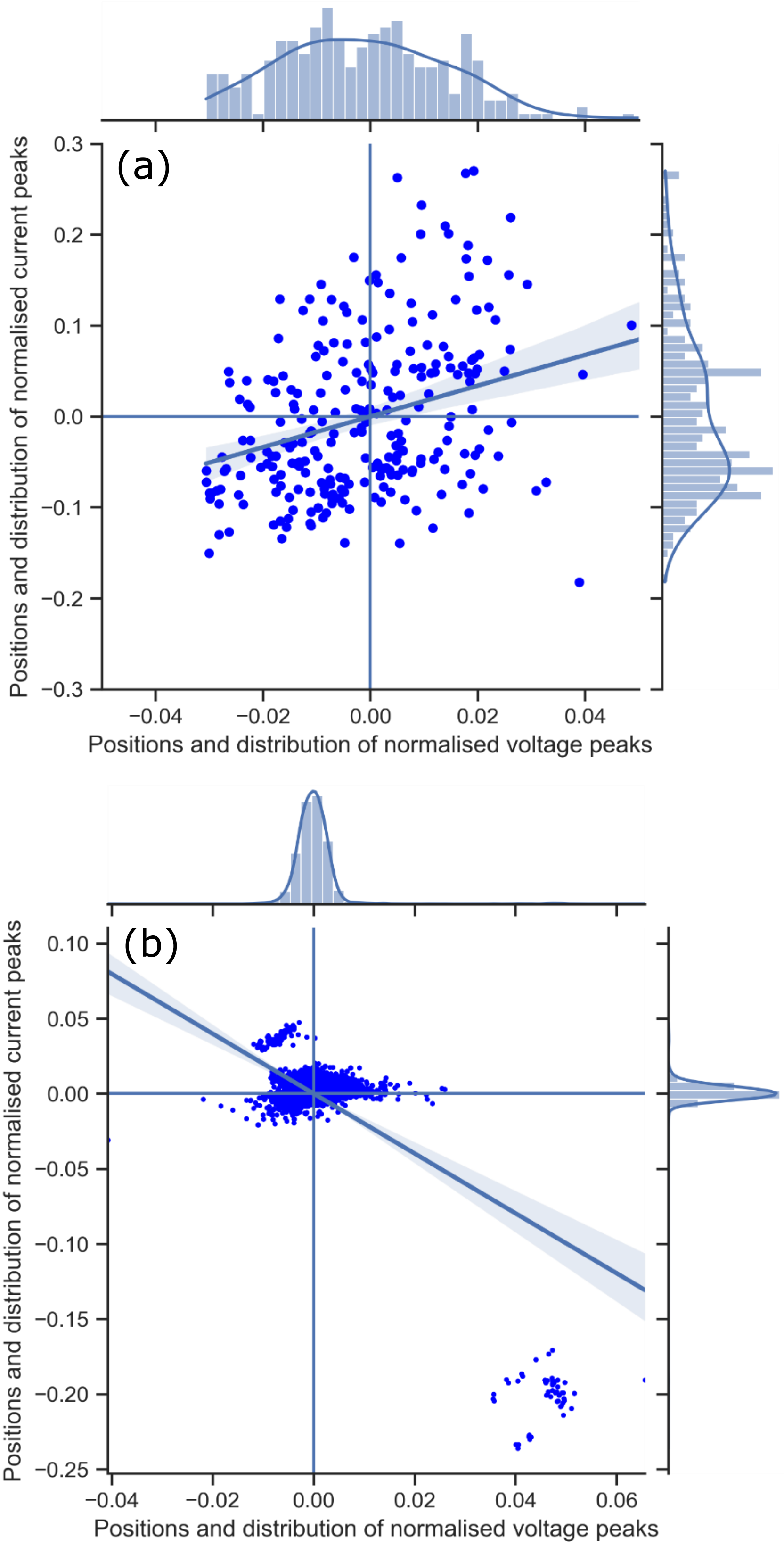

Figure 11: (a) The inter-measurement distribution and linear regression for the Gaussian fitting method. (b) The intra-measurement distribution and linear regression for the Gaussian fitting method. Note the clusters of erroneous metastable fitting at the top left, and even more noticeably at the bottom right, of the figure.

4.2.6. Moving Average Smoothing

An alternative method to minimise the effect of outliers, again by accounting for adjacent data, would be through the application of a smoothing function, such as a simple moving average. This is where each point in the data set is transformed to be a mean value derived from averaging itself with an arbitrarily sized window spanning one or either side. This transformation can be used in parallel with any method previously described but is found, for any given window size, to lead to the best performance ratios when combined with the original, maximum-finding method (although the other direct methods start to converge) and when in tandem with Gaussian fitting. The application of this method varies in computational complexity based on the window size, but is cheaper than applying a comparable least-squares function fitting, but more expensive than all methods that locate a singular data point directly.

The relationship between moving average window size and maximum performance ratio can be seen in Figure 12. In this algorithm, from a window size of zero being the same as the maximum found previously, the voltage performance ratio drops slightly. This would be due to a small increase in intra-measurement variation, in turn since at small window sizes the lower points on either side of what would otherwise be the peak would suppress the peak's altitude, resulting in neighbouring (and less suppressed) points being sometimes taken as the new peak - increasing the variation in where that peak is measured to be. After this initial reduction, however, the performance ratio increases, as the inter-measurement variation starts to gently increase, and intra-measurement variation lessens as the IV curves smooth. This relationship is maintained until a window size of around 41 (± 20 points window). At this point, while the intra-measurement variation continues to decrease the average position of each entropy element measurement start to sharply converge, dropping the inter-measurement variations and rapidly impeding the performance ratio. This leads to an optimal window size of 39 and 43, with a performance of $\sigma_r^V = 8.36$ and $\sigma_r^I = 71.6$, for voltage and current respectively. As anticipated, the mean peak values shift to the left (lower voltage), as the steeper NDR decline dominates the PDR, and downwards (lower current) since an averaged set of current values are ipso facto lower than the maximum itself. As a final note averaging 37 points in any window is quite extreme for a smoothing function, and consisting of around a third of all total data for each point. For this reason, in later comparison, a reasonable simple moving average window of 11 points will be included alongside the true optimum. This window has a performance ratio of $\sigma_r^V = 5.94; \sigma_r^I = 45.3$.

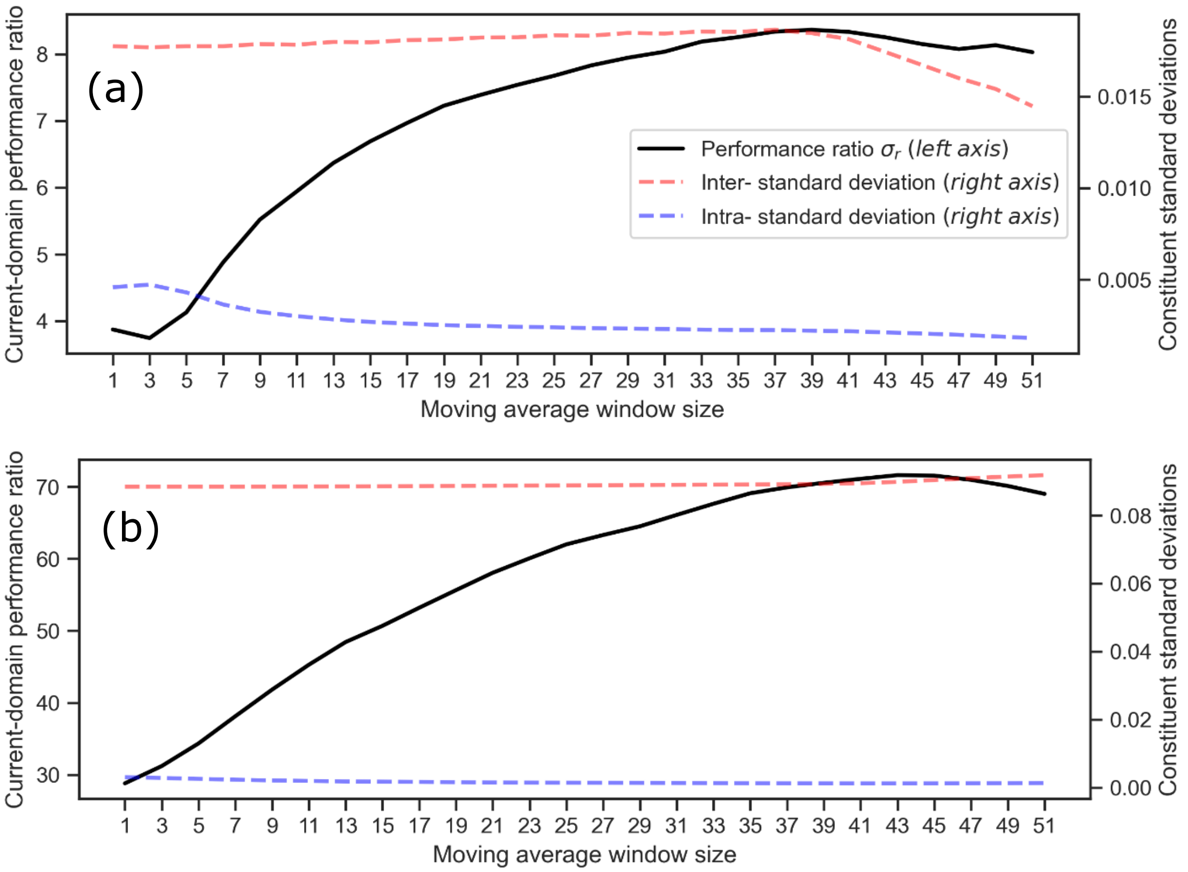

Figure 12: (a) The relationship between moving average pre-processing window size and configuration performance for the maximum method in the voltage domain. (b) The relationship between moving average pre-processing window size and configuration performance for the maximum method in the current domain. In both figures, the black line represents the performance ratio based on the left y-axis scale, while the blue and red dashed lines represent the constituent intra- and inter- standard deviations, respectively, on the right y-axis scale.

The other method meaningfully improved by pre-extraction smoothing is the Gaussian function. This method already considers multiple data points in determining the peak location, but still benefits from the additional processing resulting in a performance that exceeds taking the maximum with no smoothing. This would be due to the removal of outliers before starting as before, but also that the IV line's shape post-smoothing much more resembles the Gaussian as defined. As the averaging leads to an overall broadening of the IV line shape, the window size determined in the previous Gaussian section may not necessarily hold as optimal. Figure 13, therefore, shows the performance ratio at varying moving average window and Gaussian fitting window. As anticipated, the plot shows this increase in optimal Gaussian window as averaging window increases. The data in these plots were subjected to a 99% yield sweep extraction threshold. This is to say that the white cells in the bottom right of each figure are configurations that are unable to extract bits from more than 1% of the total sweeps in the data set – ensuring that erroneous performance isn't found as a result of the implicit null extraction of outlier values and that the method can be assured to have a negligible effect on overall bit yield.

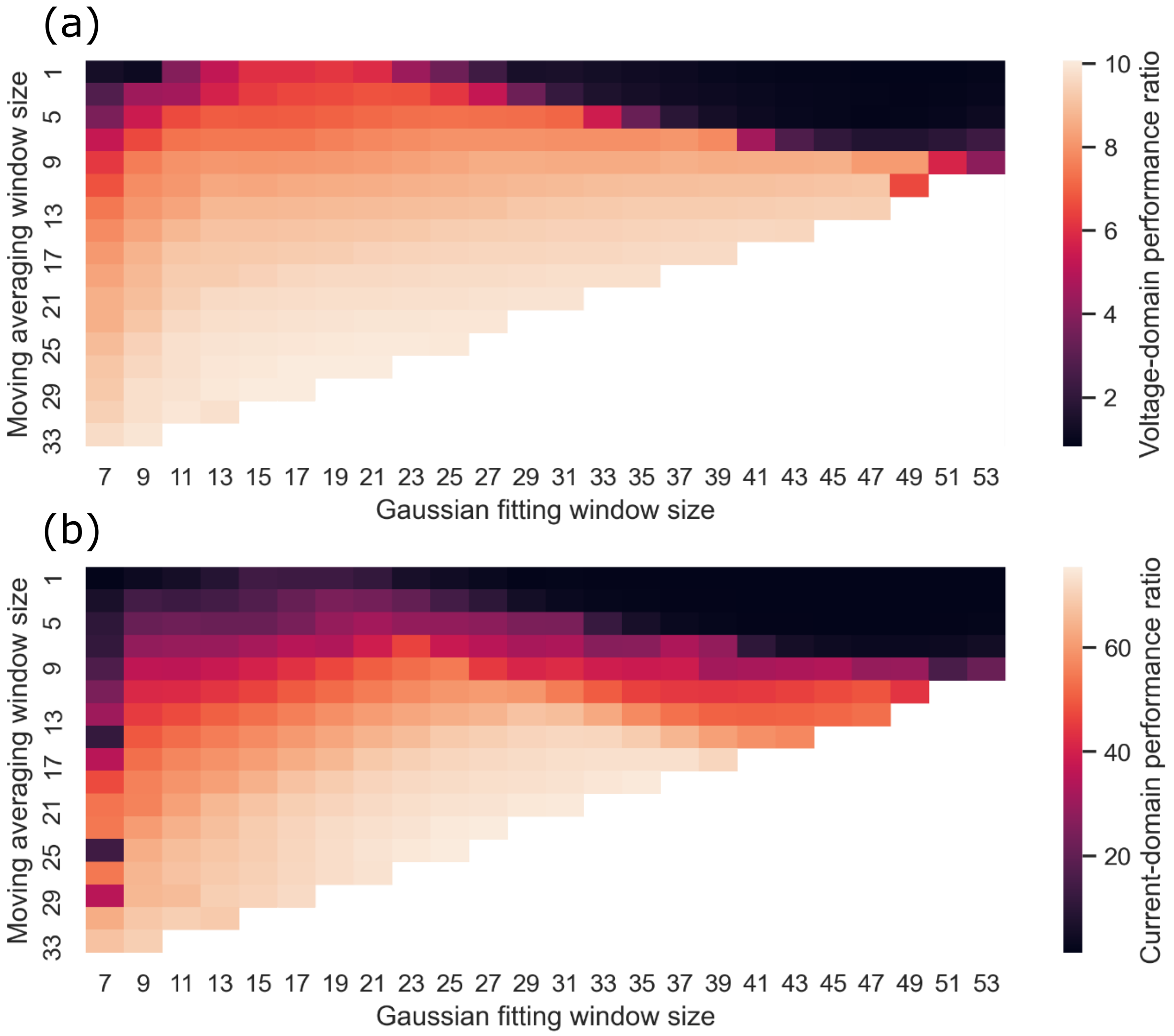

Figure 13: (a) The relationship between both moving average and fitting window size on configuration performance for the Gaussian fitting method in the voltage domain. (b) The relationship between both moving average and fitting window size on configuration performance for the Gaussian fitting method in the current domain. (Configurations that exclude over 1% of RTD measurements masked here in white)

The optimal parameters for voltage-domain performance were found to be a 27 pt averaging window and 19 pt fitting window, leading to a performance of $\sigma_r^V = 10.1$. Controlling for 11 pt maximum averaging window, as with the maximum method, gives an optimal fitting window size of 47 pt, and a corresponding performance metric of 9.15. For the current domain, the parameters were both windows were found to be optimal at 25 pt, leading to a performance of $\sigma_r^I = 75.3$ - and a fitting window of 27 pt and performance of $\sigma_r^I = 60.2$ with the 11 pt constrained averaging window. The combination of moving averaging and Gaussian fitting in the configuration employed here is found to have the most optimal extracting performance of any method that does not seek to perform additional measurements. It is also the most complex and computationally expensive, and so is not suitable for all situations, as discussed in the section directly below. In terms of repeat measurements, it will be assumed that any intra-measurement variation of the peak is derived almost entirely within the points of a single smoothing window. This means that, for instance, the lowest peak position in the natural 3σ range will be modified only by its direct centred averaging window on each repeat (with the beyond taken forward from the initial measurement), leading to an extra repeat measurement cost of half the averaging window on either side of the 3σ range. This would add to the extra measurement allocation for Gaussian fitting, also of half the fitting window on either side to ensure that the intra-measurement extrema are still treated evenly. This results in a remeasurement cost of 3σ+10=38 for the maximum technique applied to an 11-point moving average window, and a remeasurement cost of 72 and 76 points for voltage and current unbounded moving averaging and Gaussian (with a span of 21 and 24 for the fitting and averaging window sum respectively on either side).

4.3. Conclusion & Comparison

To compare these results, we must consider a few factors. The first and most important consideration is the domain in which the value measurement was taken – be it in voltage or current. On immediate evaluation, it can be assumed that it is always optimal to take a current value for the peak instead of a voltage one, as the performance ratio of current-domain extractions is around 7 times better than their voltage equivalents. This would mean that for any confidence interval there are seven times more distinct locations that an RTD's measurement can fall within, meaning that with respect to the full range of RTD values the range of a single RTD's measurement is more precise. This would have a great impact on error rates as will be seen in section 5. However, despite this sevenfold increase in performance taking current measurements has one very large drawback – unlike voltage peak position, current peak position is significantly influenced by temperature [36-38]. This makes the current method non-viable for binarisation methods that compare the value to an arbitrary constant threshold as with the most common binarisation method, featured in section 5.1, but due to the monotonic and equally-applied nature of the scaling still allow for comparative binning techniques as with section 5.2. Theoretically, an arbitrary threshold can shift with sensed temperature to maintain an appropriate position, but this added complexity is above the scope of this work. For this reason, when choosing which distributions to carry forward to the next section both current and voltage distributions will be considered, as both have differing values depending on the binarisation technique in question.

On the other hand, it can be said that while the ambient temperature can vary in a wide range, if an equilibrium is reached with ohmic heating while in operation there would a minimised effect from external temperature. This, however, adds a dependency of temperature on the power consumed by the device and would lead to non-linear shifting due to temperature in each measurement point, arising from the specifics of the current and voltage of each measurement point itself. Different philosophies as to whether to aim for more consistent, ambient-level temperatures or more externally isolated equilibrium-level temperatures for each measurement point exist, but for ease of analysis here the former approach and consequent downsides will be assumed. It is therefore also worth noting that while expected to be reasonably consistent, no measurement or control for temperature was taken in the measurement of these data. Since all repeat measurements were taken at the same time, and the different RTD measurements taken over a longer period, there is the possibility that some of the wider relative spread of RTD current values arise from unchecked variation in temperature in the first place. This could lead to an amount of erroneous inflation in the performance metric of current techniques from the data itself.

The next fundamental consideration when contrasting these techniques is the complexity of implementation. Some techniques employed in this section are very straightforward, but other techniques involve least-squares function fitting and moving averaging that require significant iterative arithmetic to employ. As a rule of thumb, the projection method and those to its left in Figure 14 can be considered as requiring trivial computational cost to extract, while the Gaussian method and methods to its right can be considered as too complex for direct onboard conversion on most simple types of hardware (with the possible exception of the 11pt MA max method). For this reason, even though a better performance can be achieved from the righter-most techniques, the most promising straightforward technique (in both domains this happens to be the maximum method) will be carried forward as well. This is also because these simpler extraction techniques require fewer data points to extract a value, meaning that more error-rectifying repeats can be applied per unit time as discussed prior.

From Figure 14, we can see that in the simple implementation collection the optimum method for value extraction was the maximum method and voltage-projection method in the voltage and current domains respectively, with the maximum method closely following for the latter as well. Due to the specific drawbacks of the projection method series, it can be argued that the small benefit over the maximum method is not worth the more systemic drawbacks in terms of pre-requisite training, the effect of outliers and the dependency on performance metric for each RTD depending on peak position. This series of studies help demonstrate that while attempts were made to establish the contrary, in terms of measurement-instance inexpensive means of feature extraction, no technique could be found to succeed over the maximum method. Second to this method in terms of measurement-instance cheap techniques is the single point hybrid method, but this is found to be optimal in a range so tight around the absolute maximum as can be considered derivative. Relatively speaking, the hybrid and decline methods offered better performance in the voltage domain over the current domain. This would be expected, for the technique of locating the peak as the value preceding a decline is less likely to vary in voltage location, but promotes the location of more extreme values in current that would trigger the process in the first place.

In comparing the more measurement-instance expensive range of methods, the first observation is that that the inclusion of moving window averaging significantly improves the performance of both the Gaussian and maximum techniques. As expected, this gain from moving average processing increases as the moving average window size is allowed to increase, and that confining the MA window to a more attainable 11 points in turn limits the gain in performance. Since the moving average window reduces the data set at a rate equal to the window size, and the Gaussian fitting can benefit from a larger fitting data set, a trade-off as to the two window sizes (and so where each data point is 'spent') exists. When this is not the case, applying moving averaging to maximum method or similar, the performance increases with average window size much further – only stopping when the points eliminated by the moving average start to encroach on the peak value range itself (at around 50 out of the 101 points). After this, it can be seen that the application of the Gaussian function outperforms the maximum function equivalent in all cases – as the methods by themselves, and at every level of moving average application. The only break from this pattern is that the unaveraged Gaussian method for current is of very notably low performance – this likely due to the technique, while trying to fit a Gaussian to the more triangular peak data, finding the least-squares solution at a greater variation in current than voltage. In other words, the Gaussian function finds that the least squares are found at a larger range of verticality compared to the horizontal position. As the data shape becomes less angular after smoothing, however, the Gaussian fit method regains its supremacy in that current domain, as with voltage. It can therefore be said that (besides measurement instance considerations) one should opt for the Gaussian fitting method over the maximum method for voltage, and in all cases should computational budget allow for averaging across the data set as well.

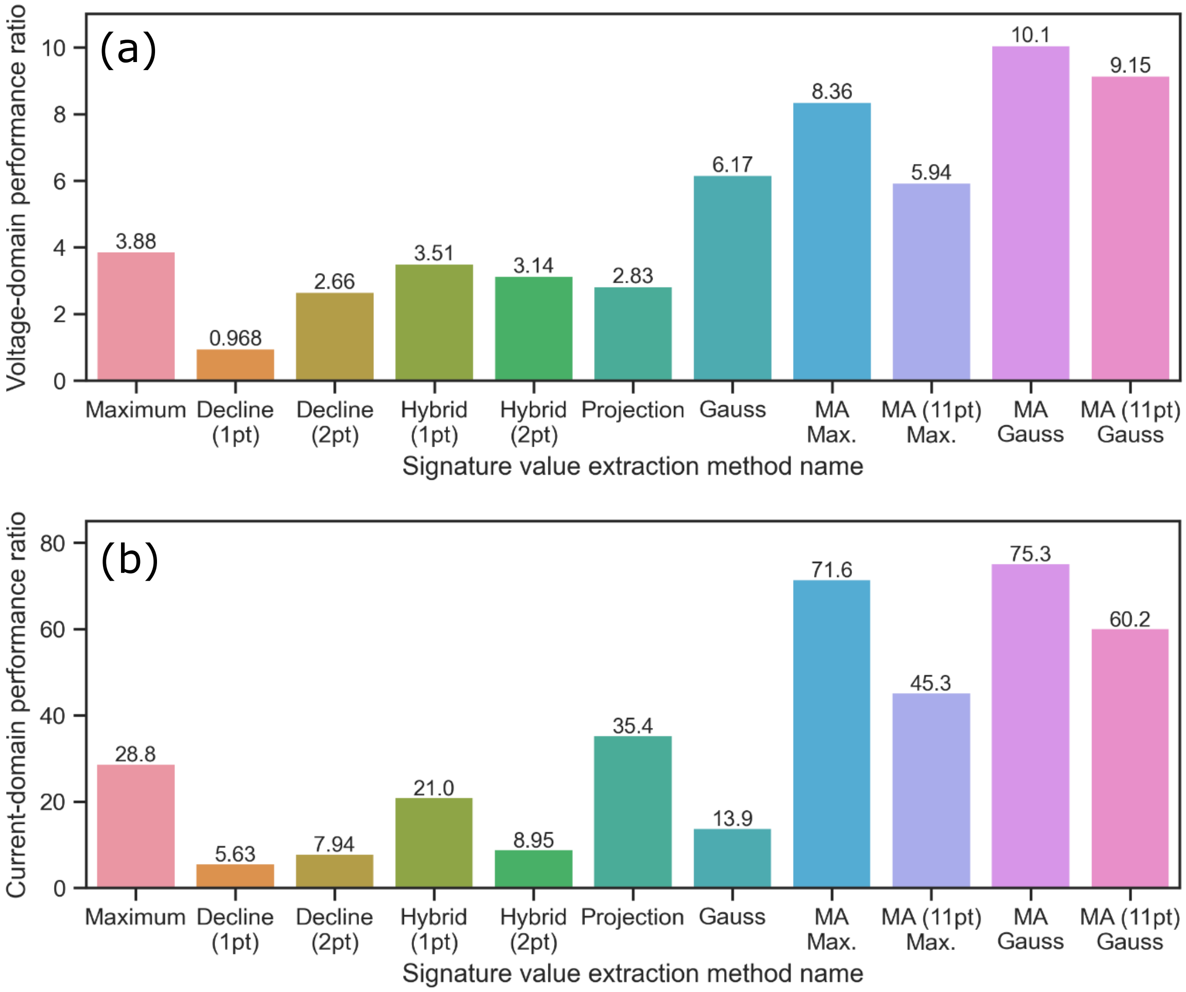

Figure 14: (a) The inter-/intra- standard deviation ratio of each technique (and select configurations) included in this study, as compared in the voltage domain. (b) The inter-/intra- standard deviation ratio of each technique (and select configurations) included in this study, as compared in the current domain. This ratio corresponds to the variation in the range of values for the full set of RTDs divided by the average variation in the measurement of each constituent single RTD of the set. In doing this, the ratio provides a metric for the distinctness of response measurements, in these figures as derived from a certain technique.

In the chapter following this one, a range of binarisation techniques are described and applied. These binarisation techniques will be applied to a more expansive data set, procedurally generated as per section 9.1 (Appendix 1), for the peak arrangements of a small selection of these extraction techniques. This will be to emulate the circumstances that would arise in a real PUF of this nature, and as such determine optimal methods and figures of merits for the stages beyond this extraction. The techniques with distributions chosen to be carried forward will be the following. First, the maximum method will be included for both voltage and current, as it was found to have the best performance metric of all non-computational methods (excluding the voltage projection on current) and is the most straightforward of any technique here considered. Next, a middle ground, requiring a middle repeat measurement requirement, for both domains was chosen – the Gaussian function for voltage and the 11 pt moving averaged maximum for current. Finally, the moving-average gauss-fitting technique (with no restraint on moving average window size) will be included for both domains, since despite being very complex and computationally expensive to calculate it boasts the highest performance metric of any method for both current and voltage.

Method	Performance ratio (Voltage)	Performance ratio (Current)	Evaluation complexity	Cost of repeat measurements
Maximum	3.88	28.8	Lower	V: 28 pt, I: 28pt
Gaussian fit	6.17		Medium	V: 42 pt, I: 38 pt
11pt MA maximum		45.2		
Averaged Gaussian	10.1	75.3	Higher	V: 72 pt, I: 76 pt

Table 5: The performance ratios, relative evaluation complexity and repeat measurement cost of the four signature extraction methods carried forward in this work

Chapter 5: Signature Binarisation

Once a value is extracted from an entropy source, such as an RTD as in the previous section, it must be converted to a binary signature to be used in a digital system. For simplicity, and to help minimise bias considerations, this is typically done by reducing each analogue element or collection of elements into a single bit state in a process here called binarisation. Various methods and protocols for signature extraction exist, but this section shall evaluate the four most fundamental. The first two sections will examine the process of binarisation by comparing an analogue value to either an arbitrary threshold or a value from a second device, while the second two will consider applying these two schemes to multiple thresholds or multiple comparisons, respectively. Within each section, this work aims to first derive an expression relating the intra- and inter- performance ratio used as the figure of merit in the previous section with the more general PUF figures of merit to be used in this chapter. As each data bit generated in the simulations and theory of this work is inherently independent, we can assume perfect unpredictability. The two important metrics are therefore the uniformity, or bias, of the system (optimal 50%) as well as the reliability, or error rate (optimal 100% or 0% respectively), of each measurement. The next section involves simulating the effect of the binning technique on a large procedurally generated data set based on both normal and KDE distributions from section 4 to determine the practical efficacy of the binning on the RTD data set and verify the theoretical expressions derived in this section, as with Monte Carlo simulation. This allows us to convert the analogue deviational ratio metrics to bias and reliability metrics for each section, which can then be compared to find the optimal binning technique for a wide variety of situations and desired outcomes.

5.1. Single Thresholding

The first method to convert an analogue signal into a binary response is to apply single threshold criteria to the value – if the value is below this threshold the value would correspond to a ‘0’ and a ‘1’ if above. This is a simple and immediate method of binarisation, taking the threshold as the mean point to generally minimise any bias that may occur. This scheme is depicted in Figure 15. It is worth noting here that basing the threshold on the *median* point of the full distribution would ensure the bias is 50%, while asymmetry or unattended outliers can easily shift a mean point threshold away from ideal. However, this median can start to be misleading if the threshold point found for one PUF is to be applied to another of the same manufacture configuration, and precludes estimation of the threshold through sampling a subset of elements on a single PUF. While it would therefore be more feasible (and employed in this work) to set generic thresholds using the mean of a distribution, it is worth noting that a more instance-tailored median could also be employed, nullifying any induced bias and corresponding diminishing of entropy. Here we also assume a perfectly average threshold, practically requiring the measurement of all elements in the set, but a study into the effects of approximating this threshold using sampling of an overall distribution can be found in external materials [43]. One could also consider a range of techniques to assess a global mean without the need for individual measurements, for example mapping a wafer of RTDs via x-ray diffraction or photoluminescence spectroscopy after epitaxial growth but before further processing.

5.1.1. Theoretical Treatment

These theoretical analysis sections seek to outline equations for the bias and error rate resultant from the figures or merit from section 4, with the assumption of the intra- and inter-measurement variables having a normal distribution. In reality, these variables would not be perfectly normal in nature (and can be seen more in terms of a skew-normal distribution in intra-measurement variation in the previous section), as found in the above section, but might be close enough to provide insight into the metrics of bias and error rate without requiring distribution-tailored numerical analysis. As an additional consideration, some PUFs such as the VIA PUF and certain concepts based around nanoelectromechanical [44] or self-assembly mechanisms [45] do not conform to a normal distribution of signature variable, and would have to be treated separately. These mechanisms typically involve negligible overlap between two groups of evaluated states (with a threshold in between), and so typically do not need such an approach for determining error rate. As mentioned previously, the most typical implementation of this binning technique is with a threshold at the estimated mean position of the underlying distribution – as can be seen in Figure 15.

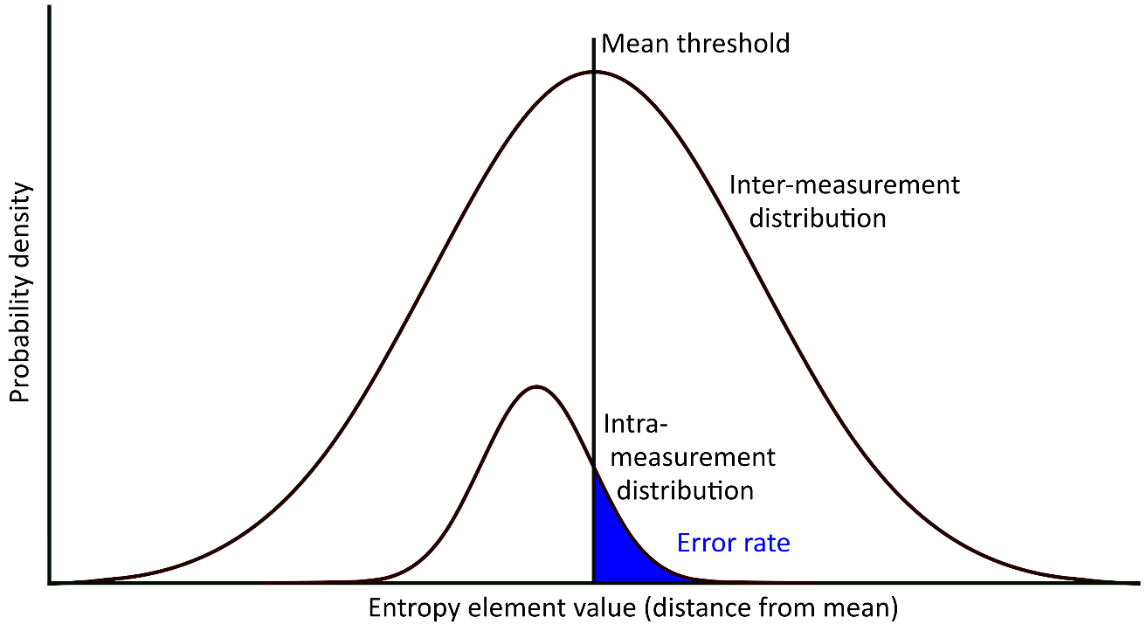
Mean threshold
Inter-measurement distribution
Intra-measurement distribution
Error rate
Probability density
Entropy element value (distance from mean)

Figure 15: Showing the single threshold at mean binning technique, in terms of the normal inter- and a single example intra-measurement distribution. The area in blue represents the probability mass of an erroneous evaluation of the entropy element (as one sample of the inter-measurement distribution, represented by the intra-measurement distribution depicted) in relation to a mean threshold.

Bias (as discussed in section 2.1.2.2.1) analysis would only look at the inter-measurement distribution, and as we are looking at the normal distribution, we know there is symmetry around the mean point. Therefore, with a threshold perfectly at the normalised inter-measurement mean $\mu_o = 0$, $Bias = F_o(0) = 0.5$, where F(x) is the normal cumulative distribution function. In other words, with these aforementioned assumptions, the single threshold at mean binning ensures an ideal bias at 50%.

The analysis for the error rate (as discussed in section 2.1.2.2.2) for the single threshold at mean is more complex to derive but is much more profound. This derivation is described in section 9.2.1, and is found to be:

$$R = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{\sigma_i}\right), \quad \varepsilon = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{\sigma_i}\right) \qquad (13)$$

Where R represents the reliability of the binning process, ε represents the error or the binning process and σ_o and σ_i represent the inter- and intra- standard deviations of the initial analogue measurement, respectively. This relatively short final equation directly relates the error rate to the performance metrics from the previous section. It offers to be a great estimator of the error rate of an analogue measurement system, based on the two measurement standard deviations and without needing numerical techniques. This relationship is depicted as the continuous line in Figure 16.

5.1.2. Experimental Results

To build up a more numerical depiction of the effects of various binning types on error rate, RTDs and measurements thereof were simulated and then binned. This was done by random sampling based on the inter- and intra- RTD distributions and allows for the verification of the expression (13) derived in the previous section. In addition, this process elucidates the effects of binning on the non-parametric kernel density estimation (KDE) as compared to the normal approximation approach. For this experimental series, the ratio of inter- and intra-standard deviation (σ_r) was varied at $0.1\sigma_r$ intervals from a ratio of 10 to (tending to) zero and the resultant error rate is taken, as can be seen in Figure 16. As this method is not resistant to value drift, the KDE of the three voltage (not including current) extraction methods carried forward from section 4 will also be included, along with the normal distributions generated from their standard-deviational parameters. For each set of inputs (sampling within normal approximated or KDE distributions), the simulation process generated 1,000 measurements of 25,600 RTDs to converge on the true value for each system.

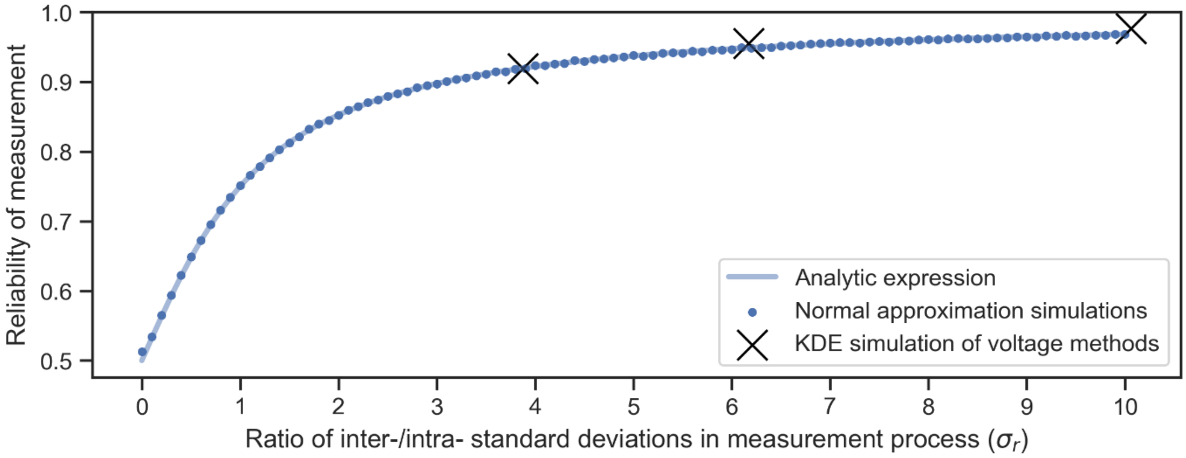

Figure 16: The relationship between inter-/intra- measurement ratio and the reliability for the single threshold at mean binning method. Blue points here represent the simulation results based on a range of normal distributions, the black crosses represent the simulation based on the KDE of the three voltage methods chosen previously (of those ratio values), and the blue line represents the expression relating the ratio to reliability as derived theoretically above.

The effects of this binning on the RTD data as extracted can be seen in Table 6 and Table 7 below.

Method name	Reliability (single threshold at mean)			
	Measurement ratio (σ_r)	Kernel Density Estimation (Non-Parametric)	Normal approximation (Parametric)	Normal approximation (Theoretical)
Maximum	3.88	91.9%	92.0%	92.0%
Gaussian	6.17	95.5%	94.9%	94.9%
MA Gaussian	10.1	97.6%	96.8%	96.8%

Table 6: The reliability of the single threshold at mean binning method for the three most attractive RTD evaluation techniques in the voltage domain. These values were taken by examining the reliability measurement as simulated numerically via KDE and normal distributions, alongside the theoretical error rates based on the equation derived in the previous section.

Method name	Bias (single threshold at mean)			
	Measurement ratio (σ_r)	Kernel Density Estimation (Non-Parametric)	Normal approximation (Parametric)	Normal approximation (Theoretical)
Maximum	3.88	46.7%	49.9%	50%
Gaussian	6.17	46.5%	49.8%	50%
MA Gaussian	10.1	47.1%	49.6%	50%

Table 7: The bias of the single threshold at mean binning method for the three most attractive RTD evaluation techniques in the voltage domain. These values were taken by examining the reliability measurement as simulated numerically via KDE and normal distributions, alongside the theoretical error rates based on the equation derived in the previous section.

From the figure and table above, it can be confirmed that the equation for error rate in the case of the single threshold at mean binarisation derived earlier fits numerical simulation and approximates the error rate in the KDE distributions, and so this analytical technique can be taken forward into section 6. The maximum, Gaussian, and smoothed Gaussian extraction techniques, with measurement ratios of 3.88, 6.17 and 10.1 respectively, have simulated reliability of above 90% in each case, with (as can be expected as per the figure) diminishing improvements as the measurement ratio increases. These reliabilities are all well above the state-indistinguishable 50% rate and can all be reasonably applied towards an implementation of a PUF without extreme compensation. In terms of bias, the KDE simulated results deviate by an average of 3.2% from the normal approximation, which allows for a reasonable approximation, but the KDE-derived bias value will be taken forward as representing the single threshold at mean binning technique for the extraction techniques included. This means that the entropy yield per bit (as defined in subsection 2.1.2.2.1) will be less than, rather than exactly, 1 for this binning technique. It may be possible to analytically estimate the bias more accurately without simulation by the proportion of inter-values above the inter-mean to below the inter-mean in the initial distribution (without the need for KDE extrapolation or normal approximation). Finally, it is worth noting again that it may not be necessary to measure the full data set from the RTDs to train a reasonable value for the mean threshold. A study into the effect of basing this mean on a limited portion of the full device set can be found in [43].

5.2. Pairwise Comparison

Alongside single threshold binarisation, the other technique commonly employed is that of single, or pairwise, comparison. This involves taking two entropy elements of the full set and comparing them, correlating the pair with a ‘0’ if, for instance, the first element was the higher value or a ‘1’ for vice versa. This is exhibited in Figure 17. For each bit to be unpredictably equiprobable, both elements of the pair can only be used once - resulting in a system with a maximum 0.5-bit entropy yield. However, assuming that the pairs are randomly chosen (or that there is no systemic gradient across devices) the bias tends to perfectly 50%, and so this reduced entropy yield decreases no further. As well as this invulnerability to bias, this comparison technique confers a further benefit. Assuming all elements are affected in the same way, this method allows for binarisation that is immune to systemic value drifting. This is to say that if all elements are equally shifting by some monotonic function, such as linearly with temperature, the higher element of a pair at one level will be the highest of the pair at any shifted level – and as such, unlike arbitrary threshold comparison, the resultant bits are independent of this change. One can imagine having an arbitrary threshold that tracks this shifting, or that enrols the distribution at multiple levels of the shifting parameter, but this would entail significant extra work and will introduce problems of its own. This has a profound impact on systems where this systemic shifting is a factor, such as with the current peak measurements of an RTD (now able to be justifiably included into these binning considerations), and for systems where entropy yield is not as important as perfect bias and strong error rate – and when post-processing improvements of bias at the cost of entropy are not preferable. A visualisation of this binning technique can be found in Figure 17.

5.2.1. Theoretical Treatment

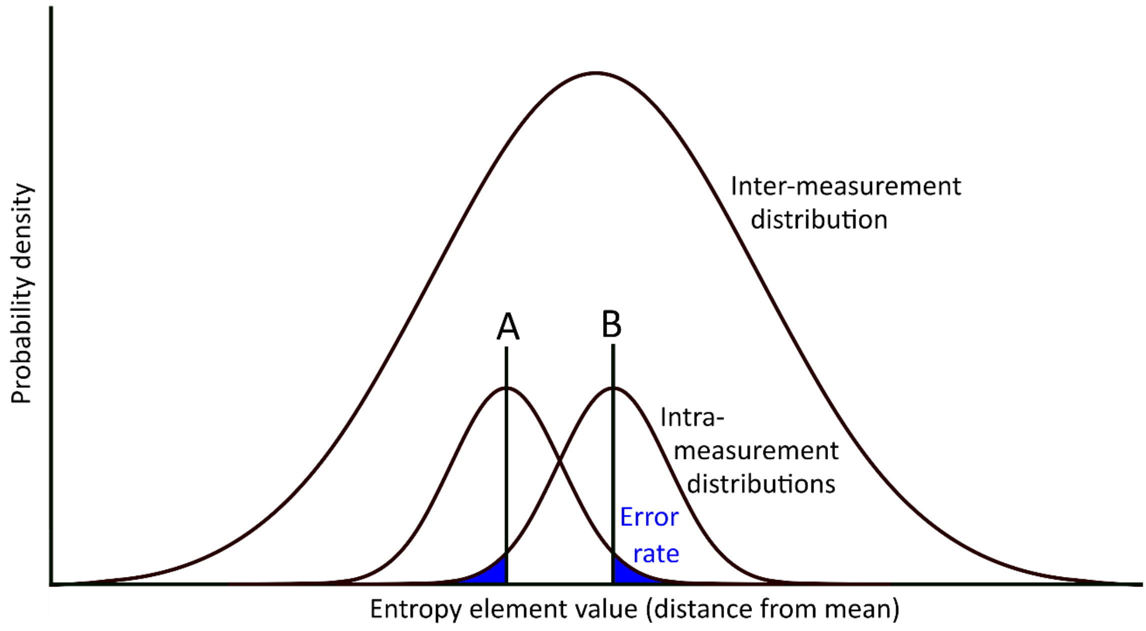
Probability density
Inter-measurement distribution
A
B
Intra-measurement distributions
Error rate
Entropy element value (distance from mean)

Figure 17: Showing the pairwise comparison binning technique, in terms of the normal inter- and two example intra-measurement distributions. The area in blue represents the probability mass of an erroneous evaluation of the order of entropy elements (as two samples of the inter-measurement distribution) represented by the intra-measurement distributions depicted.

As both A and B attend the same inter-measurement distribution, they are equally likely to be the lower or upper device respectively – and therefore the system has a bias of 50% by definition. The derivation in section 9.2.2 finds the reliability relation for this binning technique to be the same as with the single thresholding technique as Equation (13). This means, as compared to mean threshold binning, that while for the same analogue measurement ratio the reliability of the measurements will be the same, the difference in the two methods is the bias level and entropy yield. Since the single threshold method has a variable entropy yield while the pairwise comparison method is fixed at 0.5, a bias level in mean-threshold comparison can be found past which the pairwise method yields better entropy and is thus (all else being equal) optimal. Based on Equation (5) in section 2.1.2.2.1, this bias can be found to be:

$$-log_2 P_{max} = 0.5 \tag{14}$$

Where P_{max} represents the probability of the most common outcome. This means a result of $P_{max} \approx 0.707$, or a bias range of $(50 \pm 20.7)\%$. In other words, if the single-comparison method has a bias that is above a 70.7% occurrence of either bit, it is preferable to employ pairwise comparison instead. Alternatively, should threshold-training and drift resilience not be an issue, single thresholding at mean is the optimal binning choice.

5.2.2. Experimental Results

For this series of results, the normal approximation inter-/intra- deviation ratio was varied from 10.5 to $75\sigma_r$ in steps of $0.75\sigma_r$, again sampling 1,000 measurements of 25,600 RTDs. As pairwise comparison is theoretically resistant to monotonic value drift as with temperature variation, both voltage and current extraction datasets are valid. For clarity, however, only the normal approximation and the current extraction KDEs are displayed in Figure 18.

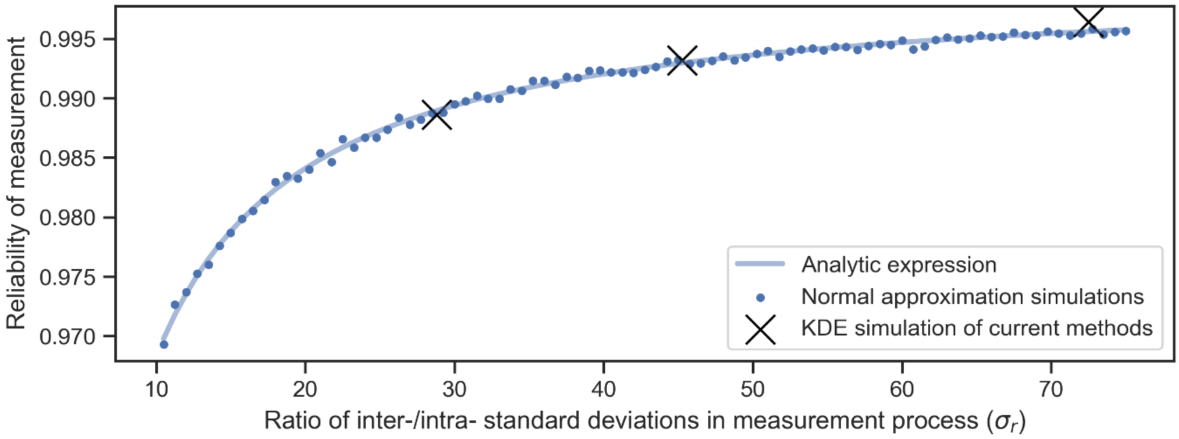

Figure 18: The relationship between inter-/intra- measurement ratio and the reliability for the pairwise comparison binning method. Blue points here represent the simulation results based on a range of normal distributions, the black crosses represent the simulation based on the KDE of the three current methods chosen previously (of those ratio values), and the blue line represents the expression relating the ratio to reliability as derived theoretically above.

The effects of this binning on the RTD data as extracted can be seen in the tables below. As all bias values were at (or extremely close and tending towards) 50%, as anticipated, the bias results are not tabulated below.

Method name	Reliability (pairwise comparison)			
	Measurement ratio (σ_r)	Kernel Density Estimation (Non-Parametric)	Normal approximation (Parametric)	Normal approximation (Theoretical)
Maximum (V)	3.88	91.5%	92.0%	92.0%
Gaussian (V)	6.17	95.2%	95.0%	94.9%
MA Gaussian (V)	10.1	97.4%	96.8%	96.8%
Maximum (I)	28.8	98.9%	98.9%	98.9%
MA Max 11PT (I)	45.2	99.3%	99.3%	99.3%
MA Gaussian (I)	75.3	99.6%	99.6%	99.6%

Table 8: The reliability of the pairwise comparison binning technique for the six most attractive RTD evaluation techniques. This is done by examining the reliability measurement as simulated numerically via KDE and normal distributions, alongside the theoretical error rates based on the equation derived in the previous section.

Again, while exhibiting a slower rate of convergence, the simulation of the binning process for normally distributed measurements follows the same expression as derived in the theoretical section above. Additionally, the KDE simulation results do not much differ from those derived from normal approximations, making it clear that the normal assumption can be made to approximate the more realistic KDE data set using the analytical expressions in the theory section above. From the trends and table, it can again be seen that there are diminishing returns in reliability as the measurement ratio increases. However, all else being equal it is of course still useful to take the highest inter-/intra-ratio extraction technique. For pairwise comparison, the entropy yield is approximately halved compared with the single threshold method, so a value judgement between footprint and reliability must be made in the process of deciding a binning implementation.

5.3. Multiple Thresholding

A distinct extension to the single threshold binarisation technique is performing comparisons against multiple thresholds of alternating bit states. This is demonstrated in Figure 19. This can also be described as direct bit extraction, as it can be considered as taking a certain significant bit out from the bit string returned by an ADC, where the bit can be considered as representative of an alternating window of a certain interval across the measurement space. This technique has the extra property of, without needing a trained mean threshold value, tending towards zero bias at the cost of increased error rate as the threshold interval shrinks. In other words, the sum of the probability density in every even interval would tend towards the sum of probability density in every odd interval as the intervals become narrower, or a less significant bit is directly taken from an ADC. In turn, a shrinking interval size comes with an increasing and narrowing of the intervals across the distribution, and as such more lines that, when crossed, correspond to the return of an erroneous bit state. This presents a trade-off between bias and error rate, found to be unfavourable as compared to a single threshold, were it not for the fact that this alternating method does not require a defined centre threshold to operate (although there would be a variation based on the offset position of the intervals in the multi-threshold implementation). This makes this method potentially more useful where finding a defined single threshold is not desirable, such as if the instance distributions vary radically between instances during manufacture, or simply to make the enrolment stage faster should the inter- and intra-measurement standard deviation ratio be strong enough to support the proportional loss in reliability as a cost.

5.3.1. Theoretical Treatment

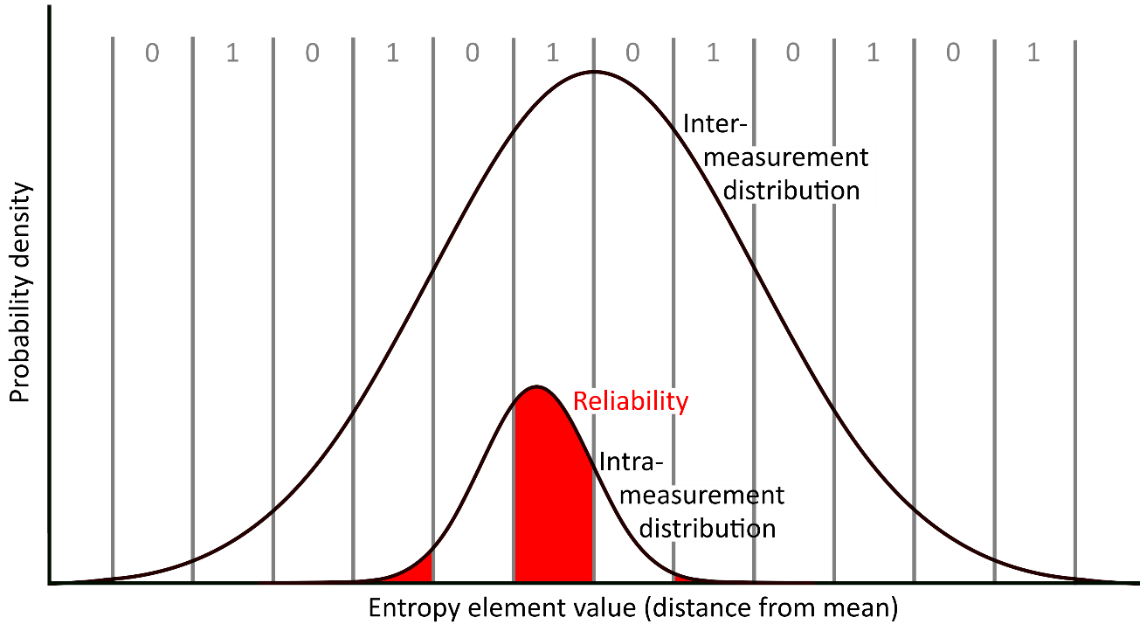
0
1
0
1
0
1
0
1
0
1
0
1
Inter-measurement distribution
Reliability
Intra-measurement distribution
Probability density
Entropy element value (distance from mean)

Figure 19: Showing the multiple threshold binning technique, in terms of the normal inter- and an example intra-measurement distribution. The area in red represents the probability mass of a correct binning allocation of the entropy element (as one sample of the inter-measurement distribution) represented by the intra-measurement distribution depicted.

The theoretical analysis for multiple thresholding lacks the symmetries of the single threshold variant to reduce to a particularly approachable expression, and includes converging series as the windows extend. A figure of this schema is found in Figure 19 and the notation is the same as can be found previously in this section. As can be found in appendix section 9.2.3, the bias of this system can be expressed as:

$$\beta_A = \frac{1}{2}(1 + \Delta\beta) \tag{15}$$

Where β_A is the bias measured for one of the bins, here denoted A, and $\Delta\beta$ represents the difference between in probability of occupation between the two states, which in turn can be expressed as:

$$\Delta\beta = \sum_{n=1}^{\infty} \text{erf}\left(\frac{2nI}{\sigma_o\sqrt{2}}\right) - \text{erf}\left(\frac{(2n-1)I}{\sigma_o\sqrt{2}}\right) \tag{16}$$

Where I represents the interval, or threshold binning width, and σ_o represents the inter-measurement standard deviation as before. This summation converges but cannot be expressed in a simpler form. An appropriate limit for the summation would be up to $n = \lceil 2\sigma_o/I \rceil$, which would account for 4σ=~99.99% of the probability mass in determining the bias (with n being factor 2, and the difference in bias decreasing further from the centre). Another practical way of setting summation limits is physical, based on the total limits of the measurement ADC (from upper to lower bound, for example from 0 to 5 V). This method would be of greater complexity but would be more directly related to the error rate. From this, we can see that as the interval reduces, or variation in measurements increases, the difference in bias between the two states decreases. The reduction in the interval can be considered as the shifting to a measurement bit one less significant, or one to the right on a typical measurement bit string. This corresponds to a halving of the interval, and the two columns within.

The reliability, derived and stated in the same appendix, is of even greater complexity. Due to the complexity of these expressions, and the high number of dependant parameters, they will not be carried forward for validation in the experimental results section. Further work, therefore, stands to numerically verify the accuracy of these statements

5.3.2. Experimental Results

Due to processing time constraints arising from the additional degree of freedom, these simulations modelled 1,000 measurements around 2,650 RTDs, an order of magnitude fewer RTDs than earlier. This series simulated these RTDs and measurements against multi-threshold binning at an interval size of 0.05 to $1\sigma_r$, in steps of $0.01\sigma_r$, where σ_r represents the ratio of inter-/intra standard deviations for each technique. The results of which, for bias and reliability, can be found in Figure 20. Here bias deviation, as the difference between the bias as found from the ideal 50%, was used. The process was applied to the normal approximation of the three voltage measurements, since this method is not value drift resistant, and just the KDE for the unsmoothed Gaussian extraction method, as the middle of the three methods in terms of complexity and prowess. The value taken for each interval was at the worst-case interval offset for all cases, at a resolution of 100 equally spaced points across the threshold interval range.

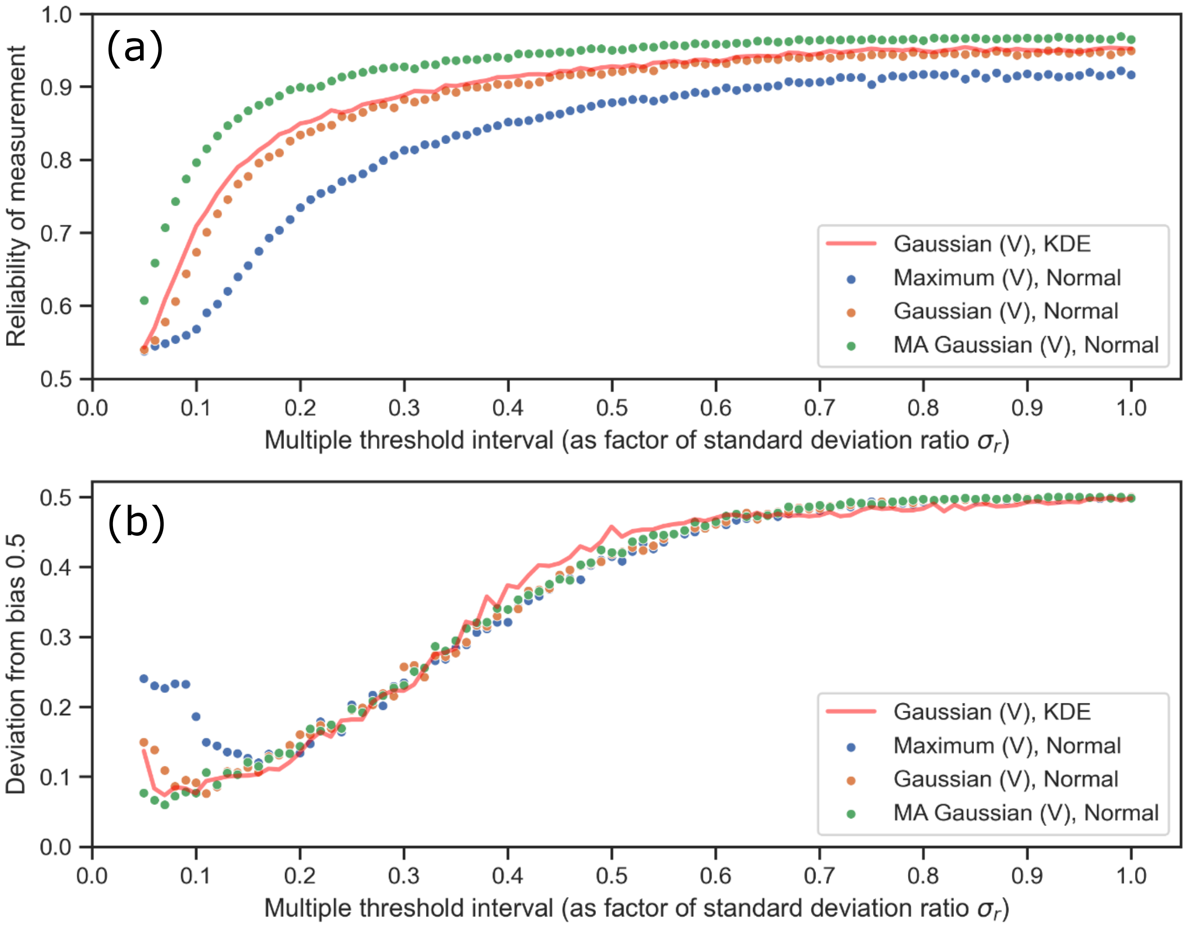

Figure 20: (a) A graph to show the relationship between the width of the multiple threshold interval and the reliability of measurements for the normal approximation of voltage maximum extraction method (blue dots), Gaussian (red dots) and averaged Gaussian (green dots). Also included are the simulation results based on the KDE for the Gaussian method – here as a red line. (b) A graph to show the relationship between the width of the multiple threshold interval and the bias of measurements for the normal approximation of voltage maximum extraction method (blue dots), Gaussian (red dots) and averaged Gaussian (green dots). Also included are the simulation results based on the KDE for the Gaussian method – here too as a red line.

From these figures, we can see that the Gaussian normal distribution roughly follows its KDE equivalent, and so can be used as a valid approximation. It is also worth noting that the simulated bias values reach a minimum at a low interval, and then start to sharply rise again, defining a limit in this domain. Since interval size can be carefully chosen in the implementation stage, the dependant variables of bias and reliability can be related to find the trade-off between the two more generally. Figure 21 shows this relationship for the normal approximations of the three voltage extraction methods.

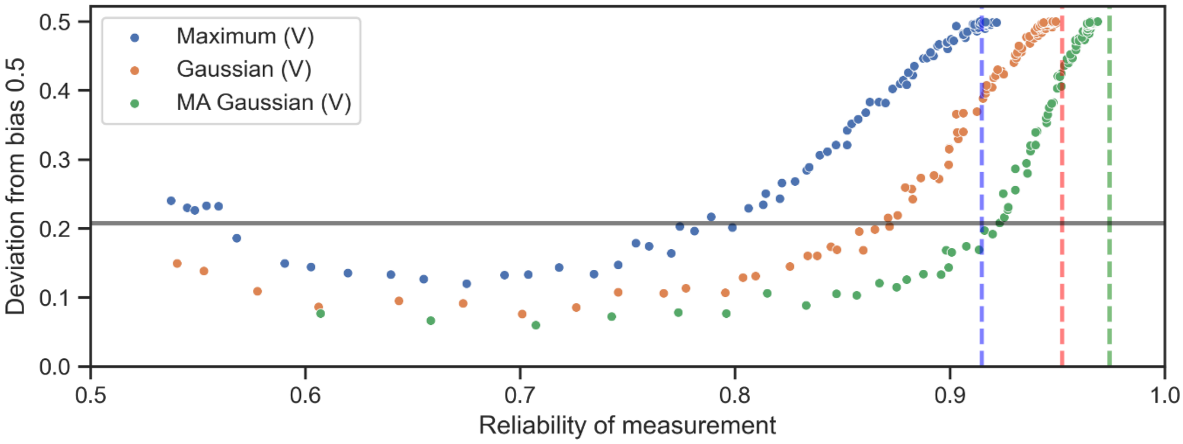

Figure 21: A graph to show the relationship between the reliability of multiple threshold evaluation against corresponding bias offset for the normal approximation of voltage maximum extraction method (blue dots), Gaussian (red dots) and averaged Gaussian (green dots). A bias with an entropy equivalent of 0.5 bits (bias offset 20.7%) expressed by a horizontal line, and the pairwise comparison technique reliabilities for these extraction methods displayed as respective vertical dashed lines.

From this, we can see that multiple thresholding does not directly outperform either the single threshold at the mean binning method (as anticipated) or the pairwise comparison method at comparable equivalent bias. Since the main feature not shared by single thresholding, this being the lack of required mean-training, is also a feature of pairwise comparison it can be said that provided an entropy of 0.5 or below is acceptable, the pairwise comparison technique is optimal to multiple thresholding. There is, however, a region between 0% and ±20.7% bias offset (1 to 0.5 bits entropy or below the horizontal line in Figure 21) that is not attainable by pairwise comparison. In this region, multi-threshold binning might be considered as the optimal binning arrangement, albeit coming at a high price in terms of reliability. Perhaps, therefore, a footprint-optimised PUF in a situation prohibiting mean-threshold-finding might find this binning technique optimal, but with a poor reliability metric and a lack of analytical certainty as to the error/bias trade-off involved caution would be advised.

5.4. Multiple Comparison

The final binarisation technique involves the concept of multiple comparisons. This would mean taking more than 2 RTDs as a subset, and converting the order of value of these elements into a single bit output, as can be seen in Figure 22. This keeps the ideal bias and immunity to drift at the cost of an error rate that increases with comparative subset size. At first glance this method appears to be a more entropically lossy version of the single comparison binning method - however, with due consideration, elements of one comparative subset may be usable in another, resulting in super-linear scaling of challenge-response pairs. This reuse cannot be done with a single pair, as too much information is leaked per bit – for instance, knowing that A is larger than B (‘1’) and B is larger than C (‘1’), then the reuse of A in a pair with C is immediately predictable as a ‘1’ as well. On the other hand, knowing that some order of A, B, C and D is associated with a ‘1‘ cannot as easily be used to help predict the bit associated with the subset A, B, C and E. This attainment of a stronger scaling comes with the same necessity of keeping the analogue elements obfuscated, in case the elements are directly measured and their value orders directly calculated, but also at the cost of vulnerability to machine learning attacks or similar on the output bits themselves. The larger the subset size, the greater the CRP space and harder (higher memory and processing requirements) the responses are to predict, at the associated cost of increasing error rate - as more of the space of possible analogue values are filled, and the chances of two similar analogue values being included in the subset increases (a further attack vector looking to derive order information from this error rate can also be imagined). Unfortunately, a meaningful study into the susceptibility of this technique to machine learning or similar attacks is beyond the scope of this work and leaves significant questions unanswered as to the feasibility of employing this method securely. The physical considerations of a side-channel attack on the source of entropy must be considered on a case-by-case basis, so is also not studied here. As a final aside, this key-value ordering must also be kept hidden from an attacker, meaning the binarisation process must be kept obfuscated, for instance on a dedicated side-channel-resistant chip. This could add limitations as to the complexity of the analogue value extraction technique that can be performed on lower-level hardware. Should this approach be valid, it would be best applicable to systems with strong inter- over intra- measurement ratios, looking for a challenge-response space larger than the entropy source would typically allow. This would also be where the measurements and binarisation of elements can be kept reasonably concealed, and applications that are either not of the security requirement to be concerned with machine learning attacks, or willing to carefully ensure a lack of vulnerability in the area.

5.4.1. Theoretical Treatment

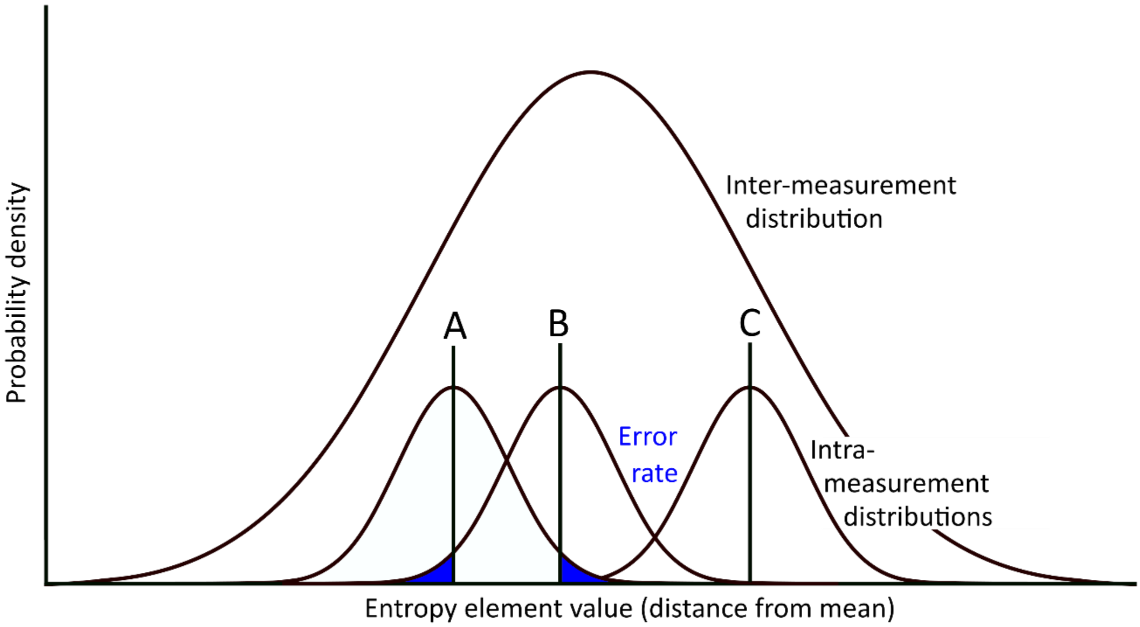
Probability density
Inter-measurement distribution
A
B
C
Error rate
Intra-measurement distributions
Entropy element value (distance from mean)

Figure 22: Showing the multiple comparison binning technique, in terms of the normal inter- and an example intra-measurement distribution. The area in blue represents the probability mass of an erroneous evaluation of the order of entropy elements (as three samples of the inter-measurement distribution) represented by the intra-measurement distributions depicted.

For the analysis of this final binning technique, as seen in Figure 22, the pairwise comparison expressions can be built upon. Bias for this technique can be set to 50% for either bit using a carefully chosen order-to-bit mapping. Since the number of distinct orders for N RTDs is N!, for any N>1 the number of orders is even (as always multiplied by 2 in the factorial). Therefore, for any N, there exists a mapping that connects each order permutation to a 0/1 bit state in a way that is equal in total. The reliability of this binning technique can be expanded from the pairwise comparison schema as seen in the following:

$$R = \left(\frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_i}{(G-1)\sigma_o}\right)\right)^{G/2} \tag{17}$$

Where G represents the number of entropy elements to be compared in one evaluation, or the group size, and where the σ terms are as before. In addition to the relationship between inter-/intra-measurement ratio and group size on reliability, it is also worth considering the relationship between group size and the total number of devices on the total number of CRPs derived. This represents the degree of scaling a PUF with this implementation would have in a range of cases. The first and most likely case for group implementation is where a group of size *G* is drawn from anywhere in the full set of potential devices *N*. This would correspond to a CRP size of:

$$CRP = \frac{N!}{G!\,(N-G)!} \tag{18}$$

Another method, of more tenuous possibility, would be to introduce an inclusion-order dependence on the comparison group – in other words making it matter which element was drawn first by, for instance, offsetting measurements by an amount dependant on this order. This would grow the CRP space more rapidly, in a relationship the same as above but without the factor *G!* in the denominator.

Alternatively, in certain cases, one might draw one member of several (optimally evenly-split) groups of quantity equal to the group size. This might be done to allow multiple measurements to easily happen in parallel, or in situations such as chaining RTDs using a series of multiplexers in a physical device [46]). In this case, the relationship between *CRP*, *N* and *G* would be:

$$CRP = \left(\frac{N}{G}\right)^{G} \tag{19}$$

5.4.2. Experimental Results

As with pairwise comparison, this binning technique is immune to uniform monotonic drift. As such the current domain RTD values can be used, and are the focus here. A comparative integer group size of 1 to 32 was taken as a varying parameter, with 2,560 RTDs simulated for a sample KDE (11-point smoothed maximum) and 25,600 RTDs for the normal approximation (with 1,000 intra-measurements samples in both cases). These values are displayed, along with the theoretically derived expected values, in Figure 23a. In this figure two sets of lines feature - one set tending towards zero reliability as group size increases, featuring the KDE and theoretical lines, and another set tending towards 50% reliability. The first case looks at the probability that a group of devices is measured to be in exact correct magnitude order whereas the second group looks at the resultant error rate assuming a random mapping of each group order to the 0/1 bit-state, as would be the most secure and hard-to-predict implementation of this binning technique. If this is the case, the reliability would tend to 50%, as even an incorrect order will work out to the correct bit state half the time ($R_{overall} = 0.5 + 0.5R_{exact}$), resulting in an unintended but still valuable gain to reliability. Figure 23b shows the relationship between group size and CRP size for a hypothetical PUF consisting of 256 elements for the device scaling rates discussed in the theory section above.

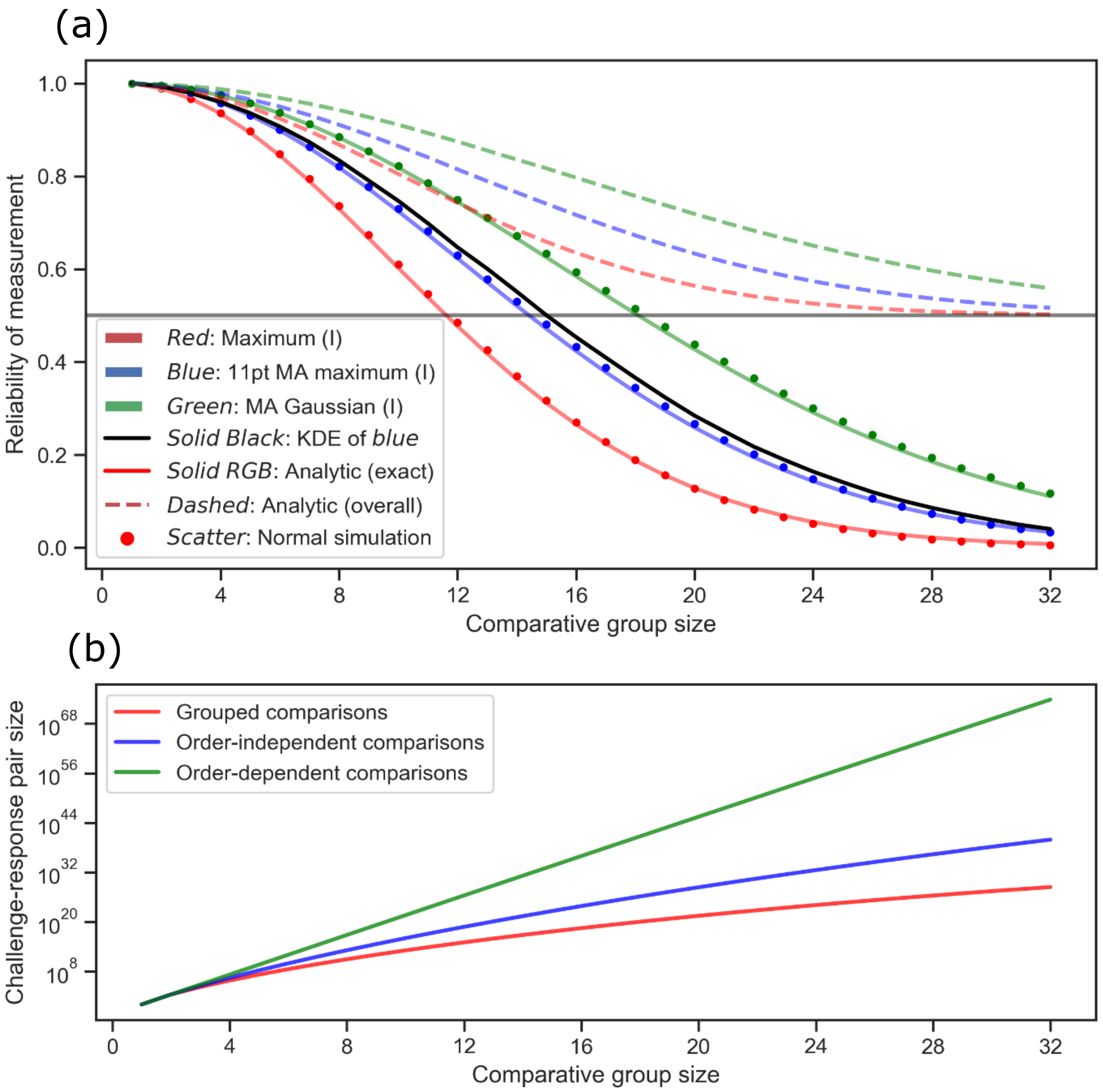

Figure 23: (a) A graph to show the relationship between comparative group size and the reliability of evaluation for the three current value extraction methods carried forward to this section. These are the maximum method (red), 11 points averaged maximum (blue) and averaged Gaussian (green). The dots of these colours represent the simulated results at these group sizes based on an approximation to normal, and the black line represents the simulated results based on the KDE for the 11pt averaged maximum extraction method. The coloured lines represent the relationship between comparative group size and error rate as derived by the analytic expression in the previous section. The solid, coloured lines represent the reliability of the exact correct ordering, while the dashed lines represent the overall bit-state reliability for a typical random-order-mapping implementation, converging to a reliability of 0 and 0.5 (black horizontal line) respectively. (b) A graph to show the relationship between comparative group size and the total CRP size for combinations of 256 RTDs, here depicted with grouped comparisons as a red line, order-independent comparisons as the blue line, and the order-dependant CRP scaling in green.)

It can be seen in Figure 23a that the KDE, simulated normal binning and theoretical normal binning all closely follow each other across group size, and as such the analytical expression for normal can be used in place of KDE simulation for any further work using this binning technique. Since the group size is independently variable, we can directly relate reliability with CRP size, as was also done with multiple thresholding and as can be seen in Figure 24. This figure relates the normal simulation results to the CRP number from 256 devices with the most probable grouping implementation of order-independent comparisons from the full set.

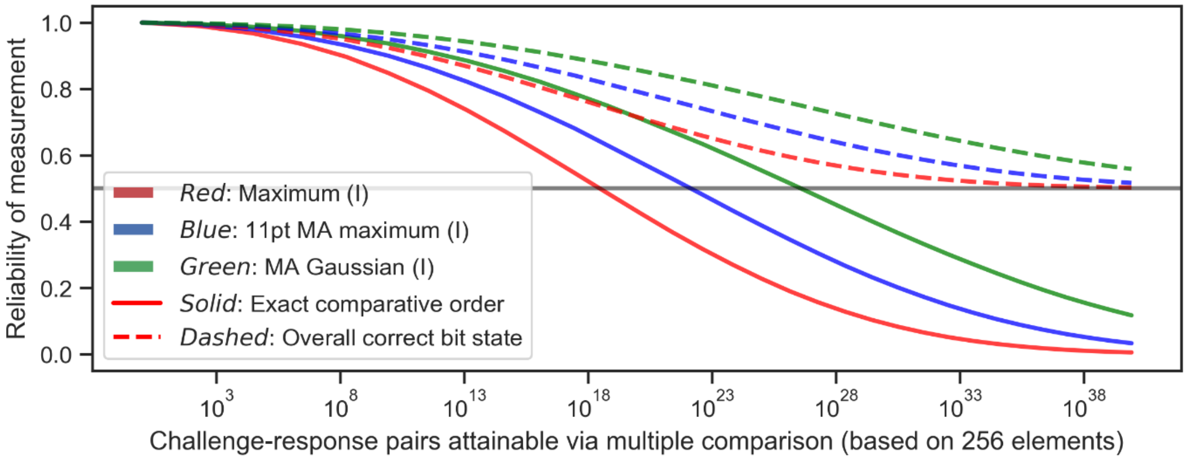

Figure 24: A graph to show the relationship between the reliability of evaluations against corresponding CRP sizes for the normal approximation of current maximum extraction method (blue line), 11-point averaged maximum (red line) and averaged Gaussian (green line). The solid lines represent the reliability of the exact correct ordering, while the dashed lines represent the overall bit-state reliability for a typical random-order-mapping implementation, converging to a reliability of 0 and 0.5 (black horizontal line) respectively. This simulation is based on groupings from a total number of 256 elements.

From this figure, we can see that a much higher CRP size is theoretically attainable with reliability that remains viable. For instance, allowing for more than 1×10^{18} CRPs to be derived while keeping the reliability above 80% in the best case (random order-bit mapping for the moving-averaged gaussian fitting RTD extraction technique in the current domain). The table below uses the analytic expression that relates group size and ratio to the reliability derived as outlined in the previous subsection, to outline the group sizes possible for each extraction performance ratio while maintaining reliability above 90%. Alongside this, similar to Figure 24, is the expected CRP space size if this implementation was applied over 256 RTDs (or other comparable elements).

Method name	Group sizes (multi-comparative binning)			
	Measurement ratio (σ_r)	Max group size for (exact chain) reliability > 90%	Max group size for (overall, rand map) reliability > 90%	CRP size for 256 devices in group size with overall R>90%
Maximum (V)	3.88	2 (91.9%)	2 (96.0%)	3.26×10^4
Gaussian (V)	6.17	2 (94.9%)	3 (92.7%)	2.76×10^6
MA Gaussian (V)	10.1	3 (90.8%)	4 (91.2%)	1.75×10^8
Maximum (I)	28.8	4 (93.5%)	6 (92.2%)	3.69×10^{11}
MA Max 11PT (I)	45.2	5 (93.2%)	8 (90.9%)	4.10×10^{14}
MA Gaussian (I)	75.3	7 (91.4%)	10 (91.2%)	2.79×10^{17}

Table 9: The group sizes possible to maintain >90% reliability with the multiple comparison binning technique for the six most attractive RTD value extraction techniques. The table includes the maximum group size for the exact order of devices and where the order is randomly mapped to bit state, alongside the maximum CRP size viable at this standard when performing the binning technique on up to 256 RTDs.

For this binning implementation it is required to achieve a reasonably large (and therefore unpredictable) group size, and therefore a high inter-/intra- performance ratio. Here, the ratios for extractions based on voltage cannot achieve a group size larger than 4, which is likely an insufficient number of parallel elements to maintain reasonable security against machine learning attacks. In the best case, while keeping reliability above 90%, 10 group elements were allowed for, resulting in a CRP space of 2.79×10^{17} from 256 devices. This is from the moving-averaged gaussian in the current domain RTD value extraction technique, with an inter-/intra performance ratio of around 75.

In comparison to an implicitly (or physically) strong-scaling PUF (maximum readout speed depending), this CRP size may be large enough to resist being sufficiently cloned by an attacker with possession of the PUF (without extra mitigation), as is the case for conventional strong-scaling PUFs. This would be a borderline case, however. To contextualise, at a read speed of 1 GHz the CRP space would be exhausted in about a decade (provided the 35 petabytes of required storage were logistically attainable) – or a significant portion, of around 1%, within around a month. Basing the binning on the second-best extraction technique, this being the 11pt-averaged maximum in current, would find the CRP space fully exhausted in the range of days. It is unlikely that a readout speed on the GHz range could be achieved, but the ideal of a strong-scaling PUF typically is many orders of magnitude above what is conceivably possible at all. However, as is the trade-off, this can be overcome by increasing the total number of devices available for comparison (as an alternative to increasing the ratio and resultant group size). For instance, doubling the device quantity (and thus footprint) to 512 increases the CRP size by a factor ~1,000, making the read time a more comfortably secure 100 years for 1% cloning at 1 GHz, all else being equal (with storage again allowing, and provided the responses can't be fully computed from a much smaller fraction of the total as with machine learning attacks).

While it is debatable whether one can comfortably credit this technique with the possessional-attack resistance of an implicitly strong PUF, there is still value in applying this scaling CRP size for redundant CRP usage against, for instance, man in the middle attacks in communication, and as one-time-pads. In the case outlined above the most effective PUF implementation of 256 RTDs could redundantly secure about 35 petabytes of data, either in storage or as part of communication. However, since the challenge (here the IDs of elements included in a comparison) consists of many bits compared to the response's one, the data stored or required for a transaction inflate greatly (provided the challenge specifics are not otherwise encoded or generated from a linear feedback shift register for instance). A simple estimation of the worst-case number of bits required in encryption would be $G \times \lceil \log_2(N) \rceil$, where G is the group size and N is the number of devices to be grouped from. For the example case of 10 devices taken from a pool of 256 as above, this would mean that every encrypted bit would require a key 80 bits long, which may become prohibitive unless otherwise encoded, and generally not competitive with conventional encryption algorithms.

5.5. Conclusion & Comparison

5.5.1. Comparison

Below is a table comparing the binning methods discussed in this work. In this table $P_{max} = 0.5 + |\text{Bias} - 0.5|$, or probability of the most probable outcome, σ_o represents inter-measurement standard deviation, σ_i represents intra-measurement standard deviation and G represents multi-comparative group size, as per the relevant subsection.

	Single Threshold	Pairwise Comparison	Multiple Threshold	Multiple Comparison
Bias (normal approx.)	50%	50%	Varies - see section	50%
Bias (w/ RTD dataset)	Simulation: Normal @ 50% KDE @ ~46.5%	Simulation: Normal @ 50% KDE @ 50%	Varies - see section	Simulation: Normal @ 50% KDE @ 50%
Entropy expression	$-log_2 P_{max}$	$-0.5\ log_2 P_{max}$	$-log_2 P_{max}$	N/A - see section

Typical entropy w/ RTDs	Simulation:
	Normal = 1
	KDE ≈ 0.902
		Simulation:
		Normal = 0.5
		KDE = 0.5
			Varies - see section	N/A - see section

Reliability expression	$\frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{\sigma_i}\right)$	$\frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{\sigma_i}\right)$	see section	$\left(\frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{(G-1)\sigma_i}\right)\right)^{G/2}$
Maximum reliability w/ RTDs	Normal @ 96.8% KDE @ 97.6% (MA Gaussian, V)	Normal @ 99.6% KDE @ 99.6% (MA Gaussian, I)	Varies - see section	Varies - see section
Mean training	Required	Not required	Not required	Not required
Drift resistance	No	Yes (uniform and monotonic)	No	Yes (uniform and monotonic)
Features	Optimal entropy yield within 50±20.7% bias, best reliability (equals pairwise comparison)	Ensures 50% bias, resists monotonic drift, best reliability (equals single threshold)	Tuneable bias/ reliability, at cost of both	CRP scales super-linearly, at cost of reliability and exposure to machine learning attack

Table 10: A comparison of the general properties of the four binning techniques presented in this work.

5.5.2. Conclusion

In conclusion, this section derived analytic expressions for, and simulated the effect of, various binning techniques on data, extrapolated both parametrically (as normal distributions) and non-parametrically (as KDEs) from the RTD data taken in the previous section. It found that the realistic KDE extrapolation simulation was similar enough to the parametric-generated simulation, which was in turn very similar to the theoretical analytic expression derived in each case. This means that the distribution based on the RTD data set can be approximated as the parametric normal and can be well described by the expressions arrayed in the comparison section. This allows the next steps in analysing the PUF process to proceed analytically rather than numerically, as seen in the next section. It is also worth pointing out here that in every analytic expression the argument was the ratio of the inter- and intra-measurement standard deviations of the initial data set, furthering the validity of this figure of merit. More generally, it has been determined that binning analogue values using a single threshold at mean is the most optimal for any given ratio of inter- and intra- measurement standard deviation, provided the data is symmetrical enough to allow for a resultant bias within 29.3% to 70.7%. This method, within this bias range, has a greater entropic yield for the same reliability than the pairwise comparison method. If this condition is not met, the same error rate but greater entropy yield can be found with the pairwise comparison binning. Pairwise comparison is also valuable to allow for measurements that (uniformly and monotonically) drift, which may allow for analogue measurements with a greater deviational ratio to be used where otherwise not possible, or make viable PUF concepts that would otherwise be prohibited due to this effect. Pairwise comparison is preferable, also, in cases where it is not logistically reasonable to train a mean value from the inter-measurement distribution. In general, one of these two binning techniques is optimal for any case presented, but in certain niche cases it may be optimal to use one of the multi-binning technique extensions. In rare cases, multi-threshold binning may be useful where mean training is not possible, but where the entropy yield or footprint is paramount and worth the higher cost paid in reliability as compared to pairwise binning. Finally, a multi-comparative technique may be employed where the measurement ratio is sufficiently high, and the benefits of a super-linear thus larger CRP set would be advantageous (such as for redundant CRP usage), in cases where it is worth the introduced risk of exposure to simulation or machine learning attacks.

Chapter 6: Systemic Considerations

As well as optimising the signature value extraction and binarisation techniques, it is valuable to optimise and tailor the PUF response generation at a higher, or more systemic, level. By this it is meant, for instance, considerations involving repeating measurements of the same entropy element to reduce error rate (section 6.1), the effect of bitstring length and continuous evaluation on the false positive and negative rates (as the incidence rate for either outcome deviate from each other, described in section 6.2), and calculating the relationship between this more nuanced understanding of error rate in terms of measurement cost in response generation – and from this concluding as to the ideal set of all the previous parameters in this work (section 6.3).

6.1. Systemic Improvements to Error Rate

This scope of consideration for this subsection seeks improvements to the error rate of the system before the stages involving, for instance, hashing and comparison with a database value. This means the error rate improvement techniques in this subsection are limited to what is considered the set of forward error correction techniques [47], which are independent of the true values. Equally, employment in PUFs prohibits (or at least greatly disincentivises) the use of error-correcting auxiliary data (for instance incremental parity bits) that would require being stored in a form of otherwise-unneeded on-board memory (in a PUF ideally the only unique element is the source of entropy, allowing for a generic or shared value extraction apparatus) and impeding the level of unpredictability for the PUF in a hard-to-analyse way. In addition, fuzzy techniques that consider an increased number of response bit configurations as permissible, typically validating based on patterns in the entropy source such as with biometric fingerprint authentication, require more than the minimum number of entropy elements to operate. This makes these techniques valuable where entropy is in abundance, such as in a fingerprint scanner, but are undesirable when each entropy element comes at a meaningful cost, as with the PUFs featured here. The final consideration here is that the fundamental ‘stored elements’ of the PUF are inherent and physically derived, and so cannot be rewritten to allow for e.g. storing convoluted, interdependent blocks instead of the bits themselves directly (damage to uniqueness parameter notwithstanding) [48]. Taking these considerations on board, the only remaining error reduction technique is through repeat measurements of the PUF elements themselves - either averaging the analogue values before binarisation or performing majority voting afterwards. These reduce the error rate of each bit, at the cost of requiring more cycles of measurement to reach. In these cases, a lack of feedback means a fixed number of repeat measurement cycles is employed. The option of repeating measurements until a match is found and confirmation sent back (known as the automatic repeat request branch of error correction), affects the false positive and false negative rates unequally and so is examined when this nuance is introduced in the next subsection, below. Additionally, in practice, this repeating may be applied selectively, towards response elements in the case that they have a higher error rate than others. Here, however, the assumption is made that each element exhibits the same error rate, as such elements can be treated interchangeably, rather than error correction being triaged towards specific, more error-prone, response elements. In this case, the relationships described below would still hold, but on the single-element level rather than overall.

6.1.1. Repeated Averaging

The first approach to error reduction with no auxiliary helper data would be to measure the analogue entropy element multiple times and then average the value before binarizing (as opposed to majority voting after binarisation, as in the next section). Taking the normal approximation validated for RTDs in section 5, we can say that the averaged-reduced error rate ε_{RA} of an element measured and averaged N times can be expressed by the function:

$$\varepsilon_{RA} = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{\sigma_i}\sqrt{N}\right) \tag{20}$$

Where σ_o and σ_i represent the inter- and intra- measurement standard deviation ratios for the distributions as before. This is derived very readily by the fact that repeated sampling of the same inter-measurement distribution (and so independent and identically distributed) reduces the standard deviation σ_i by a rate of the square root of the number of samples. From this, we can very clearly see the relationship of diminishing returns from repeat measurements on error rate. A graph of this reduction can be seen in Figure 25, taking the ratio and initial error rate of the three voltage extraction techniques and simulating the effect of repeat measurements on the error rate. To contextualise these results, it can be seen that for these voltage methods the post-averaging error rate of each method reaches the initial error rate of the more reliable result in the collection within a very small number of repetitions. However, we can also see that the repeat-averaging error rate of the direct maximum voltage method tends toward surpassing the initial error rates for the current methods (around or below 1×10^{-3}), but does not reach equality within a reasonable time (or measurement) frame (requiring over 50 repeat measurements to achieve).

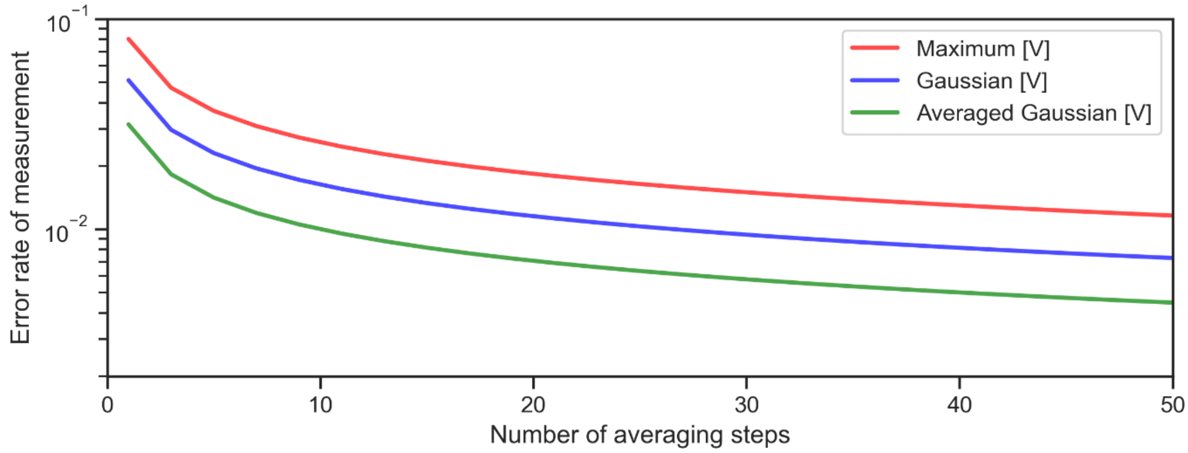

Figure 25: A graph demonstrating the relationship between the number of repeated averaging steps and resultant error rate for the three voltage-domain response extraction methods carried forward. Here the red, blue, and green lines represent the initial and subsequent error rate reductions for the maximum, Gaussian-fitting and averaged Gaussian fitting techniques respectively.

6.1.2. Repeat Binarisation

The second place where this averaging error-reduction process could occur is after the binarisation stage, examined in section 5. This would involve repeating the bit extraction and binarisation process a certain number of times, and taking the modal bit as the genuine response. Given the relative ease of the fundamental binning techniques here discussed, and that of a majority seeking process, it may not be more computationally intensive to perform repetition at this stage than managing the averaging of analogue values earlier. This earlier converting to binary processing could even be less intensive, as less floating-point arithmetic is required. The relationship between the number of repetition cycles and resultant overall error rate ε_{RB} can be expressed as:

$$\varepsilon_{RB} = \sum_{i=0}^{\lfloor N/2 \rfloor} \binom{N}{i} (1-\varepsilon)^i \varepsilon^{N-i} \tag{21}$$

Where N is the number of extracted values to apply majority voting to, taken always as odd to ensure majority, and ε represents the initial error of the entropy element measurement (taken as ε_{RA} if combined with the previous section). The improvement of the error rate as a function of the number of repeat measurements can be seen in Figure 26, alongside the relationship between averaging repetition as above. From this, we can see a much stronger error-reduction effect as compared to repeat averaging, with the error rate of the methods dropping by orders of magnitude (and surpassing any initial error rates) in a small number of repetitions. Also included, as Figure 27, is the relationship between error rate and each permutation of averaging and multiple binarising up to 50 points. From these figures, we can see that (for any number of repetitions above 2) the error rate reduces most effectively through majority voting post-binarisation, and that for any combination of either technique it is optimal to invest repeated measurements entirely into performing the multiple binarisation error reduction without the employment of averaging analogue values. In discounting the analogue averaging technique, this simplifies considerations into error reduction, and with this information we can determine the repeat measurement cost to achieve any given bit error rate. However, as discussed in the following section, one must make considerations as to bit-length, continuous evaluation (or simple automatic repeat request protocols) and the consequent splitting of error rate into the false positive and negative rate.

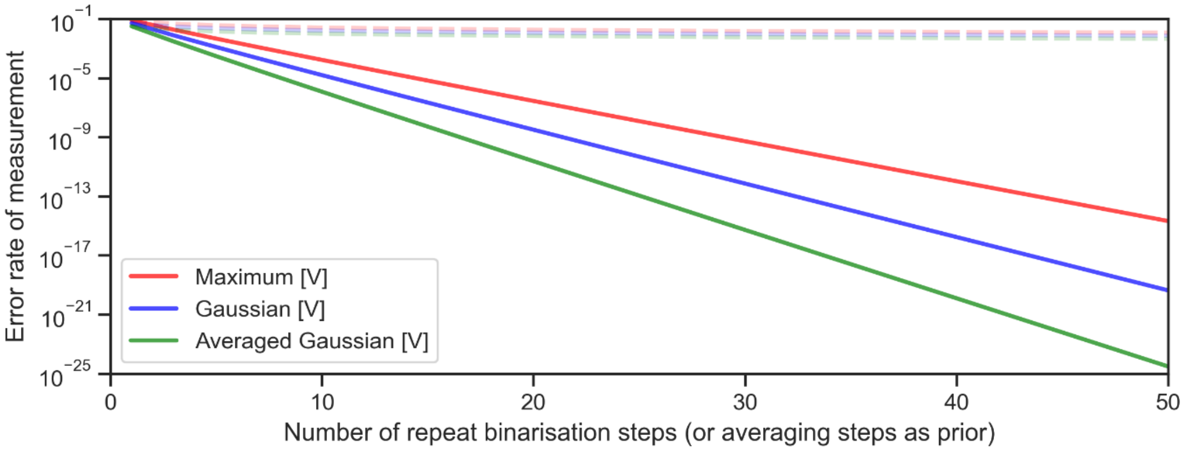

Figure 26: A graph demonstrating the relationship between the number of repeat binarisation steps and resultant error rate for the three voltage-domain response extraction methods considered in this work. Also included are the three repeat average lines as before (here translucent & dashed, at the top of the figure) for comparison. Here the red, blue, and green lines represent the initial and subsequent repeat error rate reductions for the maximum, Gaussian-fitting and averaged Gaussian fitting techniques, respectively.

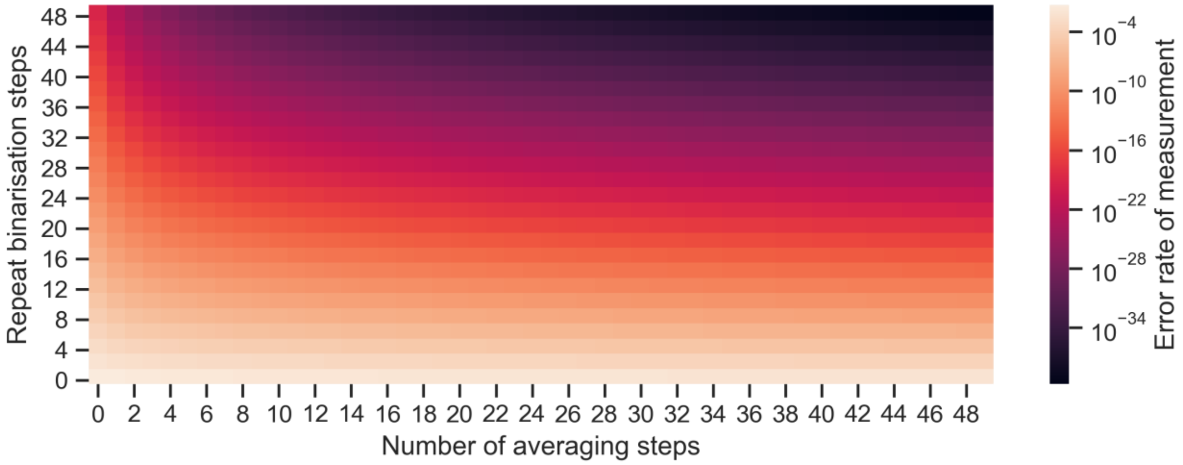

Figure 27: A three-dimensional plot comparing the resultant error rate against any combination of up to 50 repeat-averaging or repeat-binarisation steps, applying this systemic error reduction to the Gaussian-fit voltage extraction technique.

6.2. Multi-bit Systemic Considerations

As well as seeking to optimise the per-bit error rate as seen in the previous section, considerations are needed when it comes to the concatenation and comparison of full response bit strings. By itself, a single bit is insufficient to validate identity, and so multiple bits must be grouped to form the unique response. In increasing the bit-length, the possibility of a correct response being measured as incorrect (the false negative rate, or FNR) becomes more likely, while the possibility of an incorrect measurement being measured mistakenly as correct (the false positive rate, or FPR) shrinks. This splits the error rates as considered earlier into two components, and it is in terms of these two new figures of merit that the error rate is now considered as, in this stage and beyond.

6.2.1. Response Length

The first multiple bit factor consideration, separating the error rate into false positive and false negative rates, is the length of the response bitstrings, chosen to adequately assure the veracity of the PUF. This response length can be directly related to key length entropy in more conventional cryptography. The main difference between conventional keys and PUF responses is, due to the relatively higher probability of bit mischaracterising, the existence of a higher, quantifiable false positive rate in PUFs. This is compared to the more limited cases of a user input error or a very unlikely and particular digital signal noise. Increasing the PUF response bit-length decreases the false positive rate and increases the false negative rate with the following relationship:

$$FPR_{BL} = FPR^N = \varepsilon^N \tag{22}$$

$$FNR_{BL} = 1 - (1 - FNR)^N = 1 - (1 - \varepsilon)^N = 1 - R^N \tag{23}$$

Where N represents the bit-length and ε and R represents the error rate and reliability, respectively. FNR_{BL} here represents the FPR after the response length calculation process. One might immediately think that provided a response bit-length N does not drop the overall false positive rate below 0.5^K, where K is the recommended bit-length of a conventional key, it would be of comparable security to that key. This is not the case, however, as most PUF implementations would still provide an attacker with the opportunity to input a false (not entropy element derived) key with a likelihood of correctness of 50% per bit as a guess, and that each new PUF has a likelihood of any bit matching another PUF at 50% per bit randomly. This means that a suitable response bit-length minimum must be the same as is guidance for conventional keys (with the same correctness of 50% per bit), and so for the purposes of this work, therefore, a response length of 256 bits will be used. As the entropy element derived error rate must be below 50% for the valid operation of a PUF, a conventional evaluation at this response length must have a lower false positive rate than the bit-equiprobable conventional key or response-guessing method, meaning that a response length longer than the random case (here 256) is not needed for equal security either. The decrease in the false positive rate of evaluation in relation to response length can be seen in Figure 28a. As a further point of note, this high response size would correspond to a severe increase in the false negative rate, the effects of which can be seen in Figure 28b.

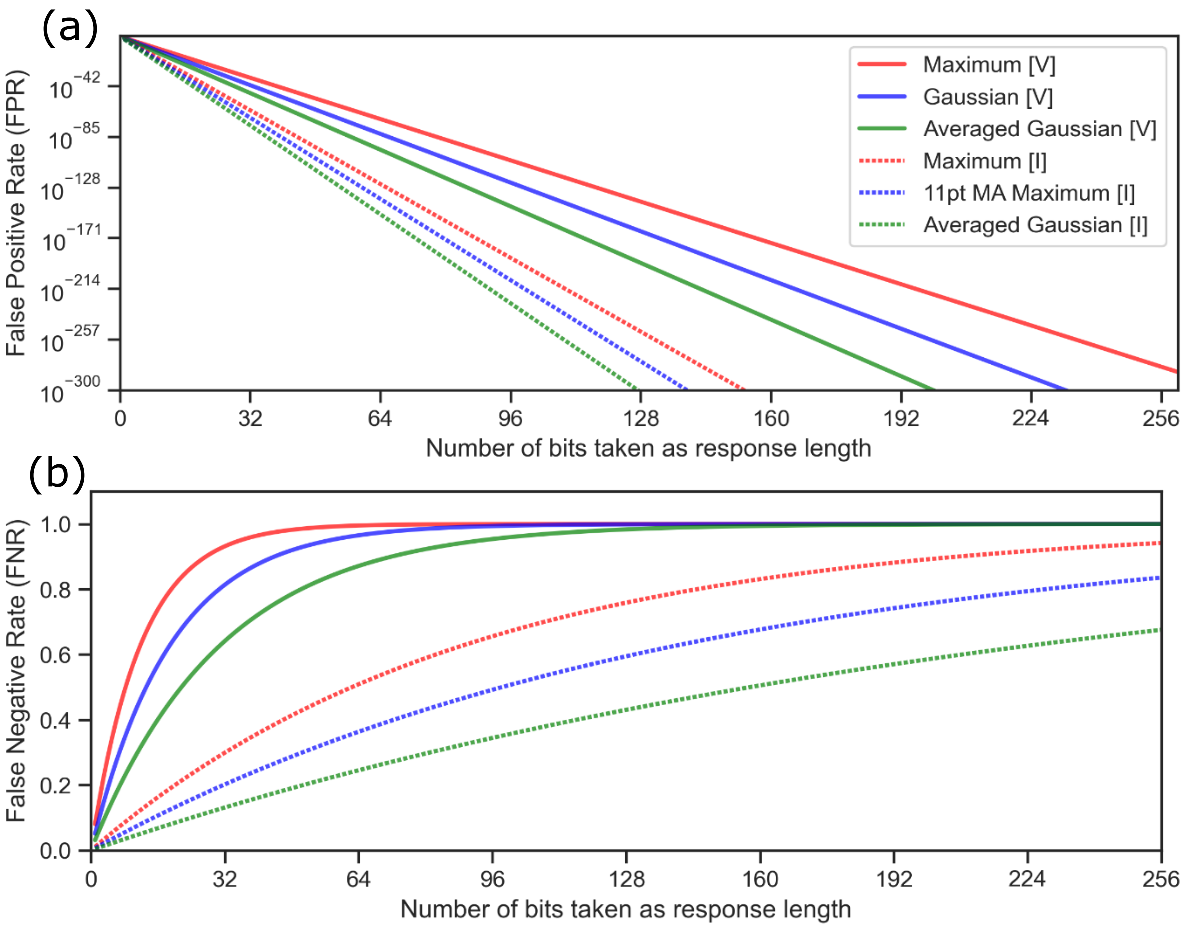

Figure 28: (a) A graph showing the reduction in false positive rate with increasing response bit-length, starting from the error rates of the 6 extraction and binarisation configurations carried forward to this section. (b) A graph showing the relationship between response bit-length and false negative rate, with the same starting errors and legend as in (a).

Finally, it is also worth noting here that this entropic view of key length security takes the equivalent bits as reduced by the (min) entropy yield rather than PUF elements themselves, and as such is adjusted by the bias or extraction yield of PUF (as introduced in section 2.1.2.2.1). In the case of the PUF case studied in this work, taking the more accurate kernel density estimation simulation of the mean threshold method in section 5.1 finds a min-entropy yield as low as around 0.9 entropic bits per PUF element (as opposed to a total yield as low as 0.5 bits per PUF elements with pairwise comparison). Whether it is necessary to increase the number of PUF elements to $256/0.9 \approx 285$ to ensure the equivalent entropy, or whether the alternative of an equivalent key length of $256 \times 0.9 \approx 230$ to keep the rounded number of PUF elements will suffice is subjective (as both can be sufficiently high). For comparison later in this section, and since a more significant entropy yield drop is induced by current pairwise comparison, all methods will be adjusted to the 256-bit key entropy equivalent, with no compromises for roundness where min-entropy yield is similar. A tabulation of the number of PUF elements, FPR, and FNR for each extraction technique can be seen in Table 11 below.

Response acquisition method	PUF elements (footprint) for 256-bit key equivalent	False Positive Rate (no systemic error correction)	False Negative Rate (no systemic error correction)
Maximum (V)	282 (@ 0.467 bias)	$\sim 10^{-281}$	$1-(5.37 \times 10^{-10})$
Gaussian (V)	285 (@ 0.465 bias)	$\sim 10^{-331}$	$1-(1.51 \times 10^{-6})$
MA Gaussian (V)	279 (@ 0.471 bias)	$\sim 10^{-383}$	$1-(2.42 \times 10^{-4})$
Maximum (I)	512 (2 RTDs per bit)	$\sim 10^{-502}$	0.941
MA Maximum (I)	512 (2 RTDs per bit)	$\sim 10^{-552}$	0.834
MA Gaussian (I)	512 (2 RTDs per bit)	$\sim 10^{-614}$	0.642

Table 11: The footprint, FPR and FNR of the six most attractive RTD evaluation methods, with optimal signature value binning.

From this table, we can see that such a high number of concatenated response bits results in a very low rate of false positive readings, but with an unreasonably high theoretical incidence of false negative events. It can be seen that even for the technique with the lowest inherent reading error, it is more probable that an in-fact correct response at this length is measured as incorrect, rather than as its true state. As we have established that for a guess-resistant secure key length FPR is not a concern, the error-derived metric of importance for the quality of the PUF would therefore be its FNR. This FNR can be reduced by improving the error rate, as outlined previously, but can also be bettered at the cost of FPR (as the converse of increasing bit-length) by repetition of the PUF response matching process in full. In other words, by repeating the bitstring comparison to database value after both bit derivation and concatenation, as in the next section. It is also worth noting here that the choice of 256-bit entropy responses over, for instance, 128 bits is dependant on application and designer tolerance. In a situation where PUF evaluation attempts are time or otherwise limited, a lower response entropy (tending towards the average human-input password equivalent entropy of 40.54 bits [49]) may be acceptable.

6.2.2. Repeat Evaluation

Now that the process of comparison to the database value has been introduced, there is one final method of forward error manipulation considered in this work, here described as repeat evaluation. This process involves performing the full evaluation and concatenation process for the PUF a number of times and comparing this repeatedly to the databased value. In the case where a validation attempt is taken as correct if any of the repeat evaluations are a match to the stored true value, then this process has the following effects on the false positive and false negative rates:

$$FPR_{RE} = 1 - (1 - FPR_{BL})^{\rho} \tag{24}$$

$$FNR_{RE} = FNR_{BL}^{\rho} \tag{25}$$

Where the exponentiation ρ represents the number of repeat evaluation cycles, FNR_{RE} represents the false negative rate after repeat evaluating, and FNR_{BL} represents the FNR at a given response length as in the previous section. Figure 29 shows this relationship between repeated measurement cycles and FPR (part a) or FNR (part b) for this technique. From this, we can see that this has the same effect as increasing the response bitstring length seen in Equations (22) and (23), only with the FPR and FNR relationships swapped – the effect here being the growth of the FPR in reducing the FNR.

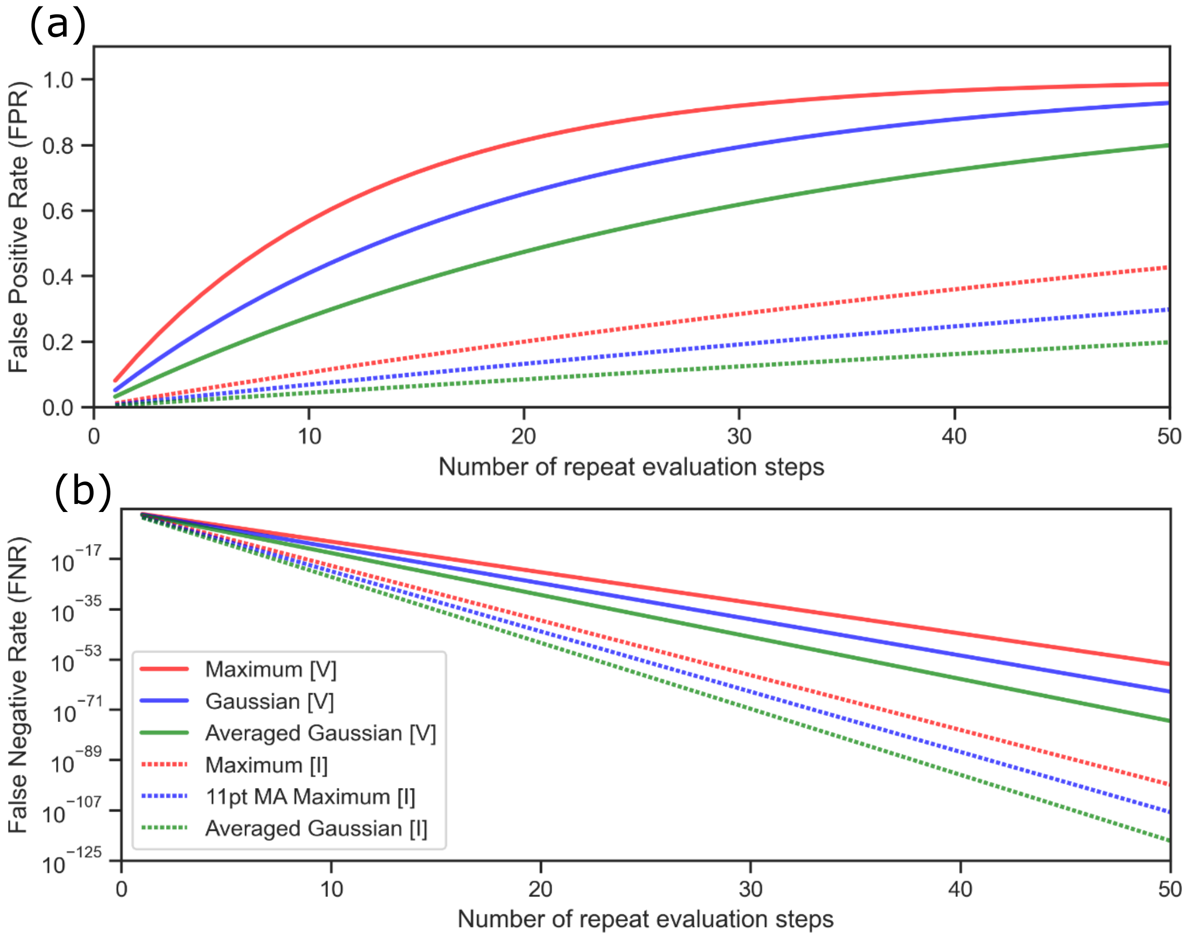

Figure 29: (a) A graph showing the increase in false positive rate with increasing repeat evaluation instances, as based on the error rates of the 6 extraction and binarisation configurations carried forward to this section. The legend for this figure is the same as and can be found in part (b). (b) A graph showing the relationship between repeat evaluation count and false negative rate, as based on the error rates of the 6 extraction and binarisation configurations carried forward to this section.

It is also possible to repeat the full comparison process several times and accept a match only if each reading is the same. This would have the opposite relationship to the any-of-set technique above (and the same as increasing bit-length). However, since this method does not affect the entropy size of the response, and that at appropriately secure key sizes the FPR is already much larger than is necessary for conventionally comparable security (at $FNR = 0.5^N$, where N represents response length), tuning the falsity rates in this direction with this method is not of value. As another consideration for the any-of-set falsity rate tuning is that while typically the overall FPR by itself cannot drop below 0.5^N (since the initial error rate per bit cannot drop below 50%), the technique described here can reduce the rate below this level if applied too heavily. This would result in a false positive rate that is less secure (rather than erring on more secure) than conventional keys. This can be remedied by increasing the response bit-length (and thus reducing the FPR), or by applying a limit of fewer cycles for this measurement repeating. To stay secure the following inequality must be maintained:

$$
\begin{gathered}
FPR_{RE} \leq 0.5^N \\
\therefore \rho \leq \left\lfloor \frac{\ln(1-0.5^N)}{\ln(1-\varepsilon^N)} \right\rfloor
\end{gathered}
\tag{26}
$$

With the latter equation derived via substitution with (24), where N stands for the bit-length of the response, ε stands for error rate as before, and $\ln(x)$ represents the natural logarithm of x. However, with any reasonable error rate ε and response bit-length N = 256, the number of repeat evaluations needed to reduce the PUF FPR to below the conventional cryptography equivalent standard (0.5^{256}) is large enough to not be a concern. So long as this adequate FPR level is sated (and so focus remains on FNR) this repeat evaluation step can be directly compared to the current frontrunner for error rate (and thus FNR) reduction – the repeat binarisation technique. The expression for FNR as a result of repeat binarisation, response bit-length and repeat evaluations can be written as:

$$
FNR_{overall} = \left(1 - \left(1 - \sum_{i=0}^{\lfloor N/2 \rfloor} \binom{N}{i} (1-\varepsilon)^i \varepsilon^{N-i}\right)^B\right)^\rho \tag{27}
$$

Where ε represents the initial error rate, N represents the number of repeat binarisation steps, B represents the bit-length and ρ represents the number of repeat evaluations. Through numerical analysis, it can be found that for any reasonable bit-length B (or, for the RTD data used in this work, starting from B = 6), there was always a stronger improvement to FNR from repeat binarisation than repeat evaluations (assuming the minimum of 2 repeats needed for majority voting). In other words, it was found that no combination employing both techniques outperformed the case where all constituent repetitions were 'spent' on repeat binarisation alone, in the same way as with repeat averaging previously. This is because even a small improvement on reliability from the repeat binarisation stage is magnified when raised to (or, physically, applied to all bits of) the response length. The repeat evaluation stage, however, occurs after the response length concatenation and can be considered as seeking to reduce the error rate after the fact. An additional downside to this method is that repeatedly checking the response against an enrolled (and externally stored) value would be associated with a higher time cost per repetition than on-chip repeat binarisation, as more instances of communication between the PUF and the database are introduced.

On the other hand, one benefit of this method is that it can be performed continuously, in other words a returned failed read can instigate a repeat evaluation until a positive match is supplied, some security- or user-experience-derived limit is reached, or the user directly intervenes. As calculated earlier, since the PUF response length would need to be as large as conventional security keys, a PUF with a 256-bit response would find limiting in this way unnecessary. In this way, this method can be employed so that the probability of a true positive converges to 1, albeit at a rate slower than fixed methods (but can of course complete faster than average if a match is found before the equivalent fully fixed error reduction technique). Finally, as an extension one can imagine an implementation that combines repeat binarisation and repeat evaluation techniques. This would be where full responses, built from the current majority bit-state of each element, are matched against the enrolled response while further repeat binarisation measurement points build (and continuously hone the reliability of each bit). This would have a stronger impact on FNR than either method alone, at the cost of additional complexity and processing time, in particular for the extended database communication as mentioned earlier.

6.3. Complete Optimisation

Now that error reduction techniques have been outlined and compared, it is possible to manipulate the overall PUF FNR (as relating to repeat measurement instances) for the extraction and binning methods examined in this work. This allows each method to be brought to the same standard of FNR and so directly compared in terms of required repeat measurement requirements. This process is performed and combined with the entropy yields found in section 5 to determine the number of entropy element (here RTD) evaluations per bit required to attain a defined security standard in each case. Once this is completed, this subsection will proceed to combine this required number of evaluations with the measurement instance (or DAC/ADC reading) cost per evaluation found in section 4 to find which extraction/binning implementation would be most efficient to evaluate to a certain security standard in a final analysis. This methodology brings together the otherwise disparate qualities of entropy yield (for example from bias), error rate and evaluation measurement cost for each PUF configuration into a single comparative metric. Unlike earlier subsections in this chapter, this subsection involves full specificity to the resonant tunnel diode PUF case study carried through this work, but the process itself is generic to the source of entropy.

6.3.1. Characterisation Instances

First, the number of entropy elements required for each PUF implementation was calculated based on the number of bits required for the desired response key (here 256) and increased to account for the loss of entropy due to measurement bias where applicable. The entropy loss accommodated for was calculated as the min-entropy, with the note that this is the theoretical best entropy yield for a certain bias, and perhaps not the yield of practical implementation. Equally, the requirement of two entropy elements per entropy bit for pairwise comparison methods (section 5.2) was also factored in. These results are considered as the footprint (entropy elements to achieve a certain key length of entropy) as relating to bit density in section 2.1.2.3.1 and are shown in Table 12. After the number of entropy elements required to extract these 256 bits of entropy was calculated, the number of times these elements must be evaluated to achieve a certain FNR was derived. This process accounted for the increase in false negative rate due to the length of the response (section 6.2.1) and assumes all error correction performed is of the most effective multiple binarisation type (section 6.1.2). Figure 30a shows the relationship between total measurement cycles (or characterisation instances, that is to say repetitions of the signature value extraction and binarisation process) and false negative level. Figure 30b shows an isolated comparison at a false negative rate of 0.1%, or 1 false negative in 1,000 matching PUF evaluations. It is worth noting that since the repetitions must occur (and consequently FNR must decrease) in discrete steps, interpolation is used to hold each technique to the same standard, resulting in non-integer characterisation instance totals.

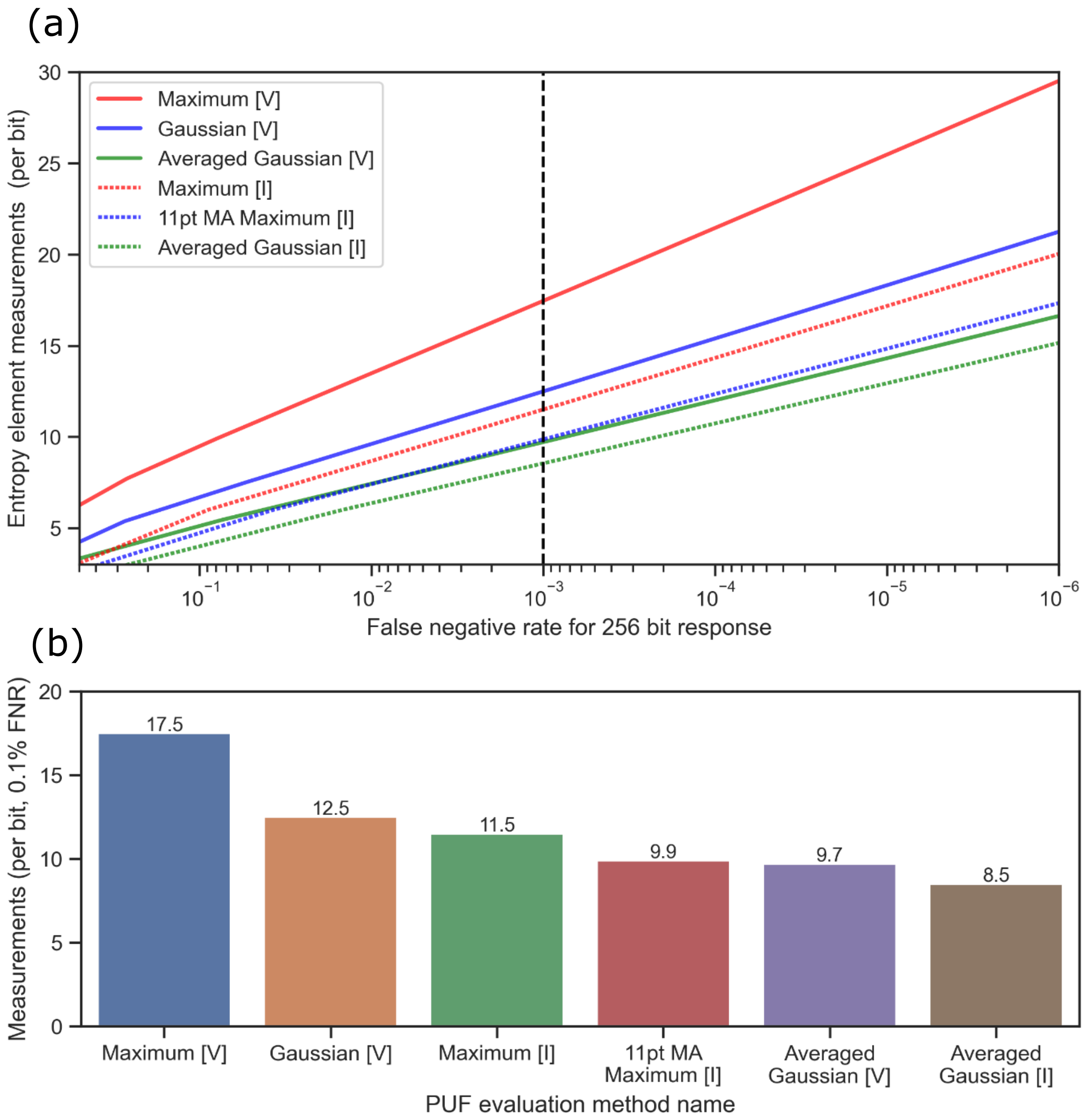

Figure 30: (a) The relationship between FNR standard and the number of full measurement cycles (or repeat signature value extractions and binarisations) per entropic bit required to achieve the standard, for the 6 extraction and binarisation techniques carried forward to this stage. The FNR standard is based on a response length of 256 bits. (b) A comparison of the 6 potential evaluation techniques compared in terms of the number of measurement cycles (or repeat signature value extractions and binarisations) per bit of entropy required to achieve an FNR of 0.1% with a response length of 256 bits.

From Figure 30a, we can see that as expected the relationship between measurement cycle and FNR is exponential (noting the log-linear axis) and at a rate that varies between evaluation methods. From Figure 30b we can more easily see that, in terms of required full extraction cycles, the moving-averaged Gaussian methods are the most efficient, with the moving-average Gaussian technique in voltage and current each requiring the lowest number of extraction instances to attain a given false negative rate in their respective domains. After these, the other two current-derived evaluation techniques attain a given FNR level most efficiently, with the moving average maximum beating out the maximum only method. This means the reduction in initial error rate (or increase in inter-/intra deviational ratio as in section 4.2) that arises from examining current over voltage overcomes the approximately double number of entropic elements required per entropy bit to fully evaluate each time. Last are the remaining techniques of Gaussian and maximum extraction of voltage respectively, in anticipated order, with the maximum voltage method requiring a significantly larger number of measurements per bit to achieve the same FNR standard. It is worth noting here that the acceptable standard of FNR that the PUF should be held to (as traded off with the evaluation time) is subjective, and a reduced measurement cycle requirement of 6.4 is required for a 1% FNR instead, for the optimal (averaged Gaussian in current) technique.

As this series of experiments is based on a response bit-length of 256, the measurement cycles required for the PUF overall are these per-bit requirements multiplied by 256. This means that the total number of peak extractions required to attain an FNR of 0.1% optimally employs the moving average Gaussian in current extraction and evaluation techniques and requires around 2,176 evaluation-level steps (1,638 at 1% FNR) to derive a response. However, this metric does not capture the full image of the relative measurement costs needed to achieve a certain error rate. As outlined in section 4, each peak value extraction technique requires a different number of data points (and so ADC cycles) to derive repeat measurements of the value taken forward for that entropy element. This means, for instance, that it takes a lot more ADC cycles, and thus time, to repeat measure the moving-average Gaussian as compared to the simple maximum, and so the subsequent reduction in error (or at that stage improvement in inter-/intra-measurement ratio) seen here does not come without cost. In the final analysis, it is the evaluation time required to derive a PUF response that dictates the cost to achieve a certain FNR, and this more accurately relates to the ADC measurement instances rather than evaluation cycles (assuming that the communication time for each PUF implementation is equal and does not dominate), so it is through this parameter that the ultimate value of each implementation will be compared – in the subsection below.

6.3.2. Total Measurement Instances

While depicting the cost of PUF operation in terms of the number of full measurement cycles per entropy element allows for easy conceptualisation, the inequality in the measurement instance cost of each evaluation technique must be accounted for. Here measurement instances mean the total number of ADC cycles, rather than total measurement cycles, in turn meaning full characterisations of an element without regard to constituent data-point measurement. When finding the location of the peak initially it can be assumed that the PUF implementation will need to examine the full 101 point range, but with the knowledge that the peak would be in a similar position upon remeasuring means repeat measurement costs vary, with the most accurate and complex techniques (moving average Gaussian fitting) having around 2.5 times higher remeasurement cost than the most straightforward techniques (direct maximum extraction, with 28 points around the previous peak containing the peak value for the next measurement with a certainty of 99.9%, as opposed to 76). With these additional considerations, the relationship between FNR and measurement instances can be seen in Figure 31, with a comparison at 0.1% FNR as before.

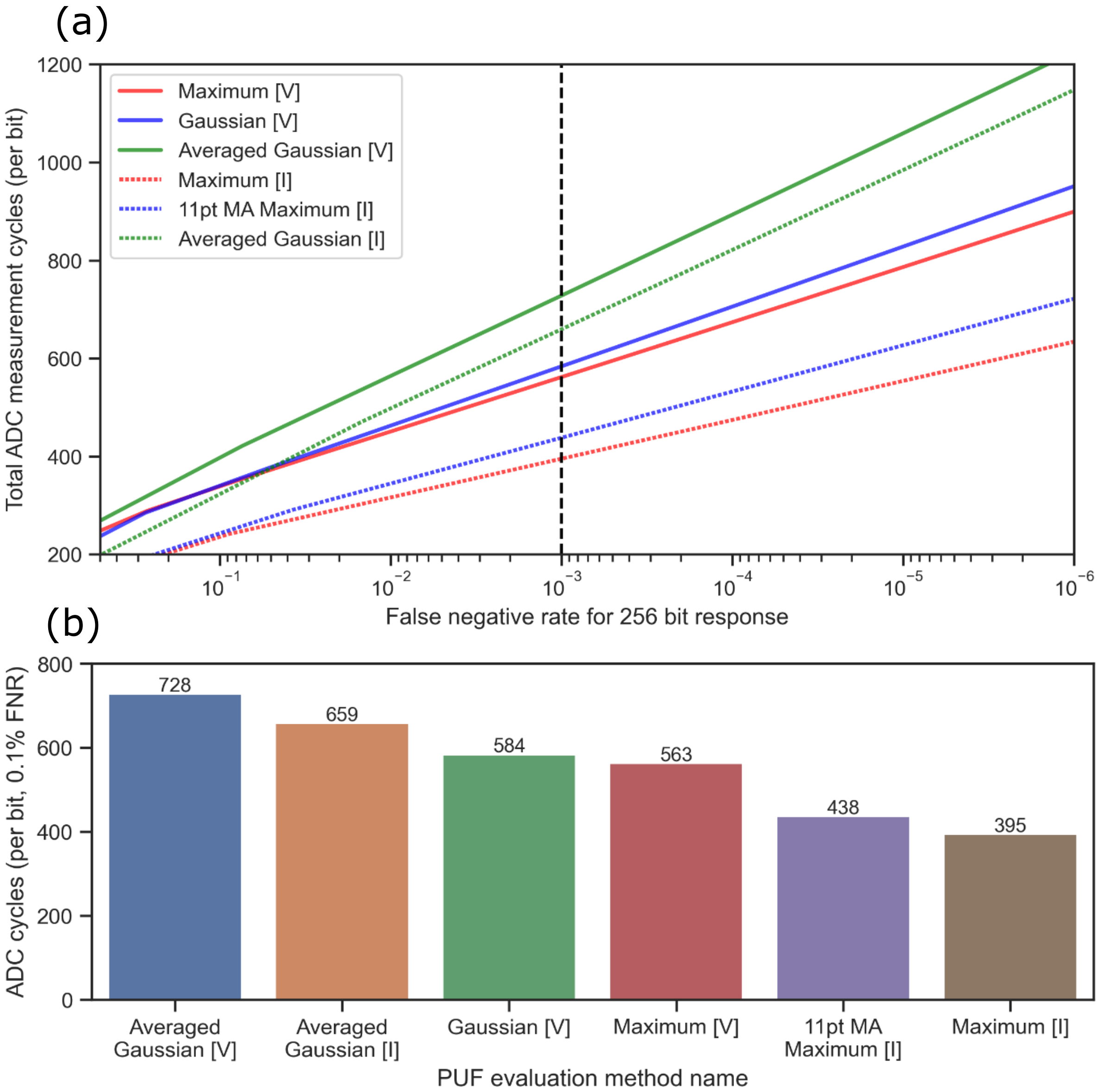

Figure 31: (a) The relationship between FNR standard and the number of ADC cycles (or RTD IV measurement points taken) per entropic bit required to achieve the standard, for the 6 extraction and binarisation techniques carried forward to this stage. The FNR standard is based on a response length of 256 bits. (b) A comparison of the 6 potential evaluation techniques compared in terms of the number of ADC cycles (or RTD IV measurement points taken) per bit of entropy required to achieve an FNR of 0.1% with a response length of 256 bits.

From this, we can see radical changes to the relative merits of the techniques as compared to measurement cycles, with different techniques being optimal at different standards of FNR. However, below an FNR standard of around 4% the evaluation techniques settle into a static order and so, given any reasonable standard for FNR would be below this value, the techniques can have a set order of value. The relationship between FNR and measurement instances is still generally exponential, as one would also expect, given the change of only a constant (initial measurement) offset and technique-differing scaling factor (repeat measurements) being different. In this representation, one can immediately see that while at the full-evaluation level (from the previous subsection) the more heavily processed moving average Gaussian fitting method requires the least cycles, the much higher number of measurement instances that constitute the method make the extraction technique the least efficient (with current methods beating out voltage domain equivalents as before). The most effective technique, requiring about half as many measurement instances per bit that the least effective voltage-averaged-Gaussian, was the direct maximum method in current. This means that despite the requirement of 2 entropy elements per response bit compared to voltage techniques and having a higher error rate than the more complex current techniques, the lower measurement instance cost of the direct maximum in the current method changed its relative standing from 4th to 1st place of the techniques here considered. To highlight this, the next most efficient method was the moving averaged maximum in the current technique, which takes the peak current value after performing a moving average with a window of 11 points – reducing the error rate by increasing the measurement instance repeat cost as a direct extension of the direct maximum method. Only after these current peak methods are the unaveraged voltage peak methods, with the maximum method the slightly more effective of the two. In other words, we can see that accounting for measurement instances per evaluation entirely flips the order of FNR efficiency in each domain, and ultimately the methods with the smaller measurement instance cost for repeat measurements beat the less error-prone but heavier competitors. As such, we can say both that in the general case a study must pay careful attention to the relative processing cost of any signature value extraction technique, and specific to this data the maximum methods (due to their quicker evaluation at this extraction level) are the optimal PUF implementation. Table 12 below captures the full results and intermediate steps of value in this process.

Signature value extraction technique	Initial error rate (& σ_r)	Entropy elements per bit (& footprint)	Evaluations/bit for 0.1% FNR @ 256-bit (& % of best)	ADC rep. cycles per evaluation (& initial)	ADC cycles/bit for 0.1% FNR @ 256-bit (& % of best)	Total ADC cycles (& meas. time @100 kS/s)
Maximum [V]	8.03% (3.88)	1.10 (282)	17.5 (205%)	28 (101)	563 (143%)	144,128 (1.44s)
Gaussian [V]	5.11% (6.17)	1.07 (274)	12.5 (146%)	42 (101)	584 (148%)	149,504 (1.50s)
Averaged Gaussian [V]	3.15% (10.1)	1.09 (279)	9.71 (114%)	72 (101)	728 (184%)	186,368 (1.86s)
Maximum [I]	1.10% (28.8)	2 (512)	11.5 (135%)	28 (101)	395 (100%)	101,120 (1.01s)
11pt MA Maximum [I]	0.703% (45.2)	2 (512)	9.86 (115%)	38 (101)	438 (111%)	112,128 (1.12s)
Averaged Gaussian [I]	0.439% (75.3)	2 (512)	8.54 (100%)	76 (101)	659 (167%)	168,704 (1.69s)

Table 12: Table of results for the analysis of the six signature evaluation techniques carried forward from section 4, as the culmination of the full process described in this work. Aside from entropy yield (entropy elements per bit, otherwise taken as 1 element per bit for voltage and 2 for current based on the symmetrical normal approximation alone), these results can be analytically derived from the initial entropy measurement properties. This process combines the otherwise disparate entities of bias and inter-/intra-measurement ratio (thus error rate) into an estimation of processing time requirements for a certain security standard without the need for the physical implementation of a device in full, along with the analytical derivation of the more typical PUF figures of merit of bias and error rate based on the initial measurements of any given normal-approximal source of entropy. Cells proportionally coloured across the red-yellow-green spectrum.

In summation, we can say that despite the entropy yield 0.5 cost for pairwise-comparison, allowing for value drift mitigations, the more precise extraction of a peak in terms of current is altogether better than the equivalent method in voltage, and overall taking the maximum point directly in current is more effective. In cases where current cannot be employed, for instance if localised heating causes non-uniform drift across the RTD elements, the optimal value extraction method in voltage is to also to take the maximum absolute value directly, rather than evaluating to fit multiple points to a Gaussian function. In light of the impact of this measurement instance cost, the projection method in subsection 4.2.4 may be revisited, despite being discounted previously due to being lossy and having consanguineous dependencies between measurement position and variation. This method has an error rate comparable to the optimal maximum methods at a measurement cost of 1 measurement instance per cycle, or 1/28th of the maximum method costs. If this reduction is worth the more holistic detriments, then with care taken this method can be studied and treated in the same way as the methods employed here. A further technique not examined here would be leveraging the inherent bistability of two RTDs in series to allow for pairwise comparison requiring only 1 measurement per pair, reducing the measurement cost to 1 measurement per bit. This would result in an error rate, arising from probabilistic switching, that cannot be determined from the datasets of individual device measurement that this work is based on, and was similarly not here evaluated.

To contextualise these results, the optimal current and overall method of finding the maximum current has a measurement cost at 0.1% FNR of 395 ADC cycles per bit, or about 1×10^5 ADC cycles to create the response overall (with the resolution as outlined in section 4). The most optimal voltage method on the other hand would require around 1.4×10^5 measurements. Assuming that the DAC communication and electronic settling time is small compared to the measuring time, and using a common general-purpose ADC sampling rate of 100 kSps, this results in a requirement of about a second of processing time for the PUF overall (with a time difference of about half a second more between the optimal current and voltage methods). To reduce this processing time, faster or parallel ADCs can be employed, the FNR standard can be reduced (for instance to an ADC requirement of 252 cycles per bit for maximum current at 1% FNR), or another adjustment to increase the implementation efficiency as above can be considered. Finally, and of most deleterious impact, the response length can be reduced – this would reduce the protection against brute force attacks, but would dampen the increase in FNR caused by the concatenation and later brought back down by increased repeat measurement cycles (and so ADC instances).

6.4. Conclusion

In conclusion, there are several methods to reduce the error rate of a certain PUF implementation and to adjust the false negative and positive rates to enhance security. This work examined the relative merits of a limited set of these methods that do not require additional data to perform, which can otherwise impede the unpredictability of the PUF response. Additionally, as a result of the high response length required to maintain security against brute force attacks, it was found that the false positive rate drops low enough not to be a concern, but also that the false negative rate becomes so high as to require significant accommodation. Of the available FNR reduction methods, it was determined that performing majority voting error correction on bits post-binarisation, rather than averaging analogue values before binarisation, repeatedly comparing full responses to the databased value, or a hybrid of the three, was optimal in reducing the error and as such false negative rate. This allowed a description of the relationship between the number of evaluation cycles (and later the number of specific measurement, or ADC cycle, instances) and false negative rate as repeat measurements occurred. The effect of bias on the min-entropy and the binarisation technique on entropy yield was then accounted for where applicable. This allowed the direct relation between the entropy yield and error rate, as FNR standard in terms of the number of repeat evaluations. This number of evaluations was then combined with the differing data-point repeat measurement requirements of each evaluation technique to ultimately relate the number of times an IV measuring must occur (in other words the number of times the ADC must operate) with a given standard of false negative rate. A comparison of the number of measurement instances required to attain a certain standard of FNR (0.1%) was made, finally providing a singular metric, in terms of ADC instance cost, to compare each evaluation technique brought forward for consideration. This comparison found that the techniques that processed the full range of measurements to get a value had a better initial error rate and were optimal in terms of evaluation cycles. However, their additional repeat measurement ADC requirements make them significantly less efficient than a more confined approach of taking the maximum point by itself or fitting a smaller window to a function in the final analysis. Discounting these data-point heavy techniques, it was uncovered that binarising the peak value in the current domain gave the best results, specifically taking the maximum value without attempting more complex processing such as applying moving average smoothing. This absolute-maximum technique was found to be optimal in the voltage domain as well, having slightly lower ADC requirements than the Gaussian technique, despite the latter's lower inherent error. Ultimately, it can therefore be said that the optimal response extraction technique in terms of evaluation time is to take the absolute-maximum-derived current value of the peak of two RTDs that binarized by comparison, requiring 512 entropy elements and approximately 1.01×10^5 measurement instances (at the resolution discussed in section 4) to achieve 0.1% FNR. Optimising for the number of required entropy elements, or if examination of the current domain is prohibited (for instance due to nonuniform drift), leads to taking the analogue value of the apex again taken directly with the maximum method. This requires an entropy element footprint of just over 256 (bias depending, in this study taken as 282) and has a measurement instance requirement of 1.44×10^5 for 0.1% FNR and the resolution taken in this work.

Chapter 7: Conclusion & Further Work

7.1. Conclusion

This work set out to define a simple, generic process to analytically derive the merits of any given physical system for employment as the entropy source for a physically unclonable function, based on the measurements of the physical system alone. In parallel with this, the process was exampled and verified for the merit derivation of a PUF consisting of resonant tunnel diodes (RTDs). These devices are non-monotonic in the current domain and operate in a manner that exploits quantum confinement to magnify the atomic-scale variation of the active region beyond that which is conventionally measurable, both enhancing the difficulty to deliberately recreate compared to similar devices. This work assumes an exact random distribution of devices, such that there is perfect uniformity, and opts for systemic techniques that both ensure that there is no detriment to holistic security and maintains an analytic nature, at the cost of optimal efficiency. However, in doing so this provides a complete process to relate a set of measurements of a physical system to an estimation of the resultant PUF efficacy.

The first step of this process was to determine the best way an entropy source, here an RTD, can be evaluated to derive a representative signature analogue value. The figure of merit derived here was the ratio of inter-measurement and intra-measurement standard deviations, which can represent how distinct each bit from one another. In other words, this is how confident one can be that any response bit measured is where it is ultimately be considered to be, in proportion to the whole distribution of points. In addition to this, the number of measurement points required for repeat evaluation was considered as, for instance, a technique that takes the apex value of the PUF’s N space IV characteristic can repeat measurements in a smaller region (based on the initial measurement) than a technique that analyses the full IV spectrum for each measurement to derive a signature value. This number of repeat measurement points has a strong effect on the overall evaluation time of the PUF. Finally, the proportion of inter-measurement values above and below the inter-measurement mean and the susceptibility of the type of value drift if any was also considered, as these relate to the footprint and viability of the PUF in later stages. From these criteria, six signature value extraction techniques were carried forward to later sections. These were three techniques of varying repeat measurement cost looking at the current domain, which has a higher inter-/intra- measurement ratio but experiences monotonic thermal drift, and three remeasurement cost equivalents from the examination of the voltage domain of the RTD, with lower measurement ratios but no significant susceptibility to drift.

The next step was the analytical relation of these properties to the PUF metrics of the error rate and the yield of entropic bits per PUF entropy source unit. These were derived analytically and verified through simulations performed on the RTD dataset. To do this, 4 fundamental techniques for signature value binarisation were outlined and compared, of which 2 can be considered as robust means of bit conversion. These were comparisons pairwise, between elements (with an entropy yield of exactly 0.5), and to a mean threshold (with an entropy yield based on bias conversion to min-entropy). Both techniques were also found to have the same relationship between the inter-/intra- standard deviation ratio to error rate post-binarisation, with the pairwise comparison technique being resistant against monotonic drift. From this, it was determined that since both binarisation techniques had equal error rate to deviational ratio efficiency, the optimal binarisation technique would be the one with the best yield of entropic bits PUF to entropy source units. Comparing the two entropy yield relationships finds that for a measurement set with a bias worse than ±20.7% the pairwise comparison technique is optimal, with the single threshold at mean method optimal when the data exhibits lower biases. Applying this to the RTD PUF development process finds that the voltage domain measurements (with biases better than ±20.7%) are best binarized with the single mean threshold technique, and while the current domain measurements are also of this better bias, they exhibit monotonic drift and must therefore be binarized by pairwise comparison at a much lower entropy yield.

The final step was to introduce systemic means of error reduction and to consider the effect of bitlength on the error rate, now split into the error rate of false positive and false negative rates (FPR/FNR). It established that for protection against brute force attacks a response bit-length of 256, as with conventional cryptography, was optimal (as each bit still had a 50% chance of being guessed with no information). This response bit-length shrank the FPR to exceptionally low, but in exchange the false negative rates for all but the lowest initial error PUFs grew to be inoperably high. This called for the employment of error reduction techniques, here chosen from options that did not expose any auxiliary data that would impede ideal predictability, and as such consisting only of repeat-processing cycles up to various stages. This also included repeat full evaluations of the PUF as a whole, which improved FNR at the cost of FPR (performing the reverse of bitlength extension). This could be performed up to a certain (albeit incredibly high for 256 bits) limit after which the FPR of the PUF based on a >50% error rate had a worse security level than a conventional key (or brute-forced attacked PUF) with its exactly 50% per-bit success rate. This section found the optimal means of forward error correction to maximally reduce false negative rates with a reasonably high bit-length and no auxiliary data was performing majority voting on repeatedly measured and binarized individual response bits, by itself in no combination with other repeat-process techniques. The required number of repeat binarisation (and so constituent signature value extractions) for a certain FNR standard was taken and combined with the repeat measurement costs and PUF elements required per bit (entropy yield) to determine the total number of measurements (or ADC cycles) required to extract a certain security level of PUF. This reduces these PUF properties to a single representative metric, and from this, the otherwise variable PUF implementation figures of merit can be directly compared. The required footprint, derived from the bit response length and entropy yield, was also considered as a final result, as minimising the number of entropy source elements in a PUF may also be a valued terminal metric. In the RTD PUF case study this step finds that the higher-complexity signature value extraction techniques, while better in initial error, are far worse than the other options in terms of the total number of measurements needed to achieve a certain FNR - and can thus be disregarded. The optimal technique was found to be extracting a bit via pairwise comparison applied to the maximum current value of the RTD's tunnelling region (with an extension to this method with greater measurement requirements and lower initial error, this being the 11-point moving average maximum method, coming in second). This technique requires the current values of the PUF to have a monotonic and universal thermal drift, but if this property cannot be ensured then a voltage domain measurement needs to be employed. Here, the optimal technique was found to be Gaussian function fitting (with single mean thresholding binarisation), followed by again taking the reading of current maximum position but in the voltage domain (the voltage position of the apex) instead. Finally, in terms of footprint voltage domain methods are all approximately of the same requirement, and require half the number of entropy elements than current domain methods. If footprint optimisation is a consideration the Gaussian voltage method is, therefore, most preferable.

7.2. Further work

There are several directions this work could take that, due to restraints on time or complexity, were not explored as part of the research found here. This section aims to outline a number of the more major venues into which further work can be taken. To expand on section 4, 'RTD Measurement', the most pertinent areas to direct attention would be looking into the employment of more directly RTD-tailored fitting functions, and the elucidation of the effect of measurement resolution on the merits of these extraction techniques. For the first of these, included in this work was the examination into the value of performing least-squares fitting the electronic measurement points to a function in extracting a signature value. This function was chosen to be a simple Gaussian due to its simplicity and the low number of free variables that can contribute to fitting metastability. However, other fitting functions may result in better feature extraction, especially given the Gaussian-fitting method was almost as efficient as the most attractive technique to extract a signature value from RTDs in the voltage domain. In particular, there would be value in evaluating fits that are derived specifically to follow the nature of the RTD IV line-shape, such as the equation below [50]:

$$J = A \cdot ln\left[\frac{1+e^{\frac{(B-C+n_1V)\varepsilon}{kT}}}{1+e^{\frac{(B-C-n_1V)\varepsilon}{kT}}}\right] \cdot \left[\frac{\pi}{2} + \tan^{-1}\left(\frac{C-n_1V}{D}\right)\right] + H \cdot \left(e^{\frac{n_2\varepsilon V}{kT}} - 1\right) \qquad (28)$$

Where J represents the current density (current per unit cross-sectional area) through the RTD and V represents the voltage across the device to produce this current. ε, k and T represent electron charge, Boltzmann constant and temperature respectively, and letters A through D, n_1 and n_2 are fitting free parameters. Overall, in this equation, the first summed term represents the current density derived from the tunnelling mechanism, and the latter term from thermionic emission (see section 2.2.1). Fitting to this function provided hopeful results in preliminary testing of single IV sweeps, but due to the complexity of the fitting process and metastability issues, this option would take significant attention to scale up to the full data set for inclusion in this study. It is likely that using functions that more closely fit the actual pattern of the RTD IV characteristic will lead to enhanced figures of merit. As a second, more important, expansion to this section it would be valuable to perform the study into extraction technique merit as seen prior with the added dimension of measurement resolution. In this work, a supplied resolution of 0.5mV in 101 points centred roughly around the tunnelling peak was taken, with a measured current resolution much higher. This resolution was taken to ensure the highest possible resolution that can be measured over the full range of RTDs and subsequent repetitions in a reasonable amount of time, rather than specifically to emulate the optimal measurement resolution of a PUF as implemented. A lower resolution data set would likely correspond to lower inter-/intra- measurement ratios, but will also reduce the number of data points required to achieve that ratio (and vice versa at an increasing resolution), which may result in a more or less optimal measurement time in the final analysis. To contextualise, for the basic 'maximum' extraction method the applied voltage resolution was in steps that were 1/10th of the intra-measurement standard deviation and 1/40th of the inter-measurement standard deviation. In the final PUF device a measurement resolution that is this high may not need to be taken, and as such the number of required ADC measurement points to achieve a certain FNR standard would reduce. This consideration could be taken for both the measurement (ADC) resolution (and as such time per ADC point, taken as 100kSps in this work), but more so the applied voltage (DAC) resolution, and may derive into a valuable general rule of thumb in terms of the measurement inter-/intra deviations.

For section 5, 'Signature Binarisation', the most notable enhancements to scope would be examining more complex binarisation techniques and further study into the multiple comparison technique. Foremost, the techniques described in this work were considered as the most straightforward principal components that would constitute more complex bit extraction techniques. This allowed the error rate to be drawn directly from the measurement deviations, but are likely not the most effective approaches and may not be the most immediate choice in certain PUF implementations. For instance, fuzzy extraction techniques borrowed from fingerprint recognition or local binary pattern (LBP) algorithms [51] borrowed from image recognition may be more valuable, with figure of merit relations that can be stated less explicitly. Related to this, it would be valuable to perform an analysis to make more tangible the predictability costs of reusing entropy elements in, for instance, the multiple comparison method discussed in the section. The LBP method mentioned earlier, as an example, in native form compares the PUF entropy elements to (a certain radius distant) neighbours in a manner that reuses each element in a given comparison more than once. Understanding the impact on the predictability of these methods with less redundant entropy element usage (with secure hashing post-binarisation considered) would help establish the feasibility of element reuse (and thus super-linear scaling) techniques as a whole.

For the error reduction stage in section 6, 'Systemic Considerations', as with the binarisation stage in main section two, the most valuable improvements to this work would be the inclusion of less theoretical-fundamental and thus more effective (if of less analytically expressible merit) techniques. Here, a study into the effectiveness of error correction techniques that require the storage of extra data, such as Bose–Chaudhuri–Hocquenghem error-correcting codes [14], can be taken and the effect of these algorithms on PUF-oriented predictability can be more explicitly discussed. Also, in expanding error correction, this work could outline a reasonable protocol for error reduction without the assumption that each PUF entropy element is identical in properties and fully independent, which may not be a justifiable assumption in practical devices. By focusing on the most errant PUF entropy units, the error rate can be reduced most effectively at the lowest measurement cost. As another point, it was found in this section that repeat binarisation was more effective by itself than combined with any separate number of instances of repeat evaluation comparison or pre-binarisation averaging. However, there may be merit in performing repeat evaluation comparisons while the majority voting or averaging build up and reduce the error themselves. This intrinsic combination of repeat evaluation comparison and the other two techniques was not studied here but stand as an interesting extension. Finally, a more formal evaluation can be performed into the anticipated physical measurement circuits to provide a more complete relation of the ADC cycle requirement to the anticipated measurement time. Certain circuit-level techniques, such as ramping voltage using an integrator capacitor or applying mean thresholding using a single comparator, may find themselves much more effective than the more digital-logic-heavy generic techniques used for measurements in this work.

Chapter 8: References

[1] U. Rührmair, J. Sölter, and F. Sehnke, "On the Foundations of Physical Unclonable Functions," *IACR Cryptology ePrint Archive,* 2009.

[2] U. Rührmair, H. Busch, and S. Katzenbeisser, "Strong PUFs: Models, Constructions, and Security Proofs," *Towards Hardware-Intrinsic Security: Foundations and Practice,* pp. 79-96, 2010.

[3] P. Tuyls, G. J. Schrijen, B. Skoric, J. v. Geloven, N. Verhaegh, and R. Wolters, "Read-Proof Hardware from Protective Coatings," in *Cryptographic Hardware and Embedded Systems (CHES)*, 2006.

[4] B. R. Anderson, R. Gunawidjaja, and H. Eilers, "Initial tamper tests of novel tamper-indicating optical physical unclonable functions," *Applied Optics,* vol. 56, no. 10, pp. 2863-2872, 2017.

[5] K. Kursawe, A. Sadeghi, D. Schellekens, B. Skoric, and P. Tuyls, "Reconfigurable Physical Unclonable Functions - Enabling technology for tamper-resistant storage," in *IEEE International Workshop on Hardware-Oriented Security and Trust (HOST)*, 2009, pp. 22-29.

[6] R. Maes, "Physically Unclonable Functions: Concept and Constructions," in *Physically Unclonable Functions: Construction, Properties and Applications*: Springer, 2013, pp. 11-48.

[7] I. Verbauwhede and R. Maes, "Physically unclonable functions: Manufacturing variability as an unclonable device identifier," in *Proceedings of the ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2011, pp. 455-460.

[8] Y. Cao *et al.*, "Optical identification using imperfections in 2D materials," *2D Materials,* vol. 4, no. 4, p. 045021, 2017.

[9] J. Roberts *et al.*, "Using Quantum Confinement to Uniquely Identify Devices," *Scientific Reports,* vol. 5, p. 16456, 2015.

[10] Y. Hori, T. Yoshida, T. Katashita, and A. Satoh, "Quantitative and Statistical Performance Evaluation of Arbiter Physical Unclonable Functions on FPGAs," in *2010 International Conference on Reconfigurable Computing and FPGAs*, 2010, pp. 298-303.

[11] A. Maiti, V. Gunreddy, and P. Schaumont, "A Systematic Method to Evaluate and Compare the Performance of Physical Unclonable Functions," in *Embedded Systems Design with FPGAs*, P. Athanas, D. Pnevmatikatos, and N. Sklavos, Eds. New York, NY: Springer New York, 2013, pp. 245-267.

[12] U. Rührmair, F. Sehnke, J. Sölter, G. Dror, S. Devadas, and J. Schmidhuber, "Modeling attacks on physical unclonable functions," in *Proceedings of the 17th ACM conference on Computer and communications security*, 2010, pp. 237-249.

[13] U. Rührmair *et al.*, "PUF Modeling Attacks on Simulated and Silicon Data," *IEEE Transactions on Information Forensics and Security,* vol. 8, no. 11, pp. 1876-1891, 2013.

[14] R. C. Bose and D. K. Ray-Chaudhuri, "On a class of error correcting binary group codes," *Information and Control,* vol. 3, no. 1, pp. 68-79, 1960/03/01/ 1960.

[15] R. Maes and I. Verbauwhede, "Physically Unclonable Functions: A Study on the State of the Art and Future Research Directions," in *Towards Hardware-Intrinsic Security*Berlin, Heidelberg: Springer, 2010, pp. 3-37.

[16] T. Mcgrath, I. E. Bagci, Z. M. Wang, U. Roedig, and R. J. Young, "A PUF taxonomy," *Applied Physics Reviews,* vol. 6, no. 1, 2019.

[17] G. Lenzini *et al.*, "Security in the shell: An optical physical unclonable function made of shells of cholesteric liquid crystals," in *IEEE Workshop on Information Forensics and Security (WIFS)*, 2017, pp. 1-6.

[18] R. Helinski, D. Acharyya, and J. Plusquellic, "A physical unclonable function defined using power distribution system equivalent resistance variations," in *Proceedings of the 46th ACM/IEEE Design Automation Conference (DAC)*, 2009, pp. 676-681.

[19] L. Wei, C. Song, Y. Liu, J. Zhang, F. Yuan, and Q. Xu, "BoardPUF: Physical Unclonable Functions for printed circuit board authentication," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2015, pp. 152-158.

[20] Y. Yao, M. Kim, J. Li, I. L. Markov, and F. Koushanfar, "ClockPUF: Physical Unclonable Functions based on clock networks," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2013, pp. 422-427.

[21] Q. Chen, G. Csaba, P. Lugli, U. Schlichtmann, and U. Rührmair, "The Bistable Ring PUF: A new architecture for strong Physical Unclonable Functions," in *IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, 2011, pp. 134-141.

[22] J. W. Lee, D. Lim, B. Gassend, G. E. Suh, M. v. Dijk, and S. Devadas, "A technique to build a secret key in integrated circuits for identification and authentication applications," in *Symposium on VLSI Circuits. Digest of Technical Papers*, 2004, pp. 176-179.

[23] J. H. Anderson, "A PUF design for secure FPGA-based embedded systems," in *15th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2010, pp. 1-6.

[24] B. Gassend, D. Clarke, M. v. Dijk, and S. Devadas, "Silicon physical random functions," in *Proceedings of the 9th ACM conference on Computer and communications security*, 2002, pp. 148-160.

[25] J. Guajardo, S. S. Kumar, G.-J. Schrijen, and P. Tuyls, "FPGA Intrinsic PUFs and Their Use for IP Protection," in *Proceedings of the 9th international workshop on Cryptographic Hardware and Embedded Systems*, 2007, pp. 63-80.

[26] F. Tehranipoor, N. Karimian, K. Xiao, and J. A. Chandy, "DRAM based Intrinsic Physical Unclonable Functions for System Level Security," in *Proceedings of the 25th edition on Great Lakes Symposium on VLSI*, 2015, pp. 15-20.

[27] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," *Solid-State Electronics,* vol. 125, pp. 25-38, 2016.

[28] L. Zhang, X. Fong, C. Chang, Z. H. Kong, and K. Roy, "Feasibility study of emerging non-volatile memory based physical unclonable functions," in *IEEE 6th International Memory Workshop (IMW)*, 2014, pp. 1-4.

[29] P. Koeberl, K. Ü, and A. Sadeghi, "Memristor PUFs: A new generation of memory-based Physically Unclonable Functions," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2013, pp. 428-431.

[30] L. Zhang, X. Fong, C. Chang, Z. H. Kong, and K. Roy, "Highly reliable memory-based Physical Unclonable Function using Spin-Transfer Torque MRAM," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2014, pp. 2169-2172.

[31] L. Zhang, Z. H. Kong, and C. Chang, "PCKGen: A Phase Change Memory based cryptographic key generator," in *IEEE International Symposium on Circuits and Systems*, 2013, pp. 1444-1447.

[32] D. Jeon, J. H. Baek, D. K. Kim, and B. Choi, "Towards Zero Bit-Error-Rate Physical Unclonable Function: Mismatch-Based vs. Physical-Based Approaches in Standard CMOS Technology," in *Euromicro Conference on Digital System Design (DSD)*, 2015, pp. 407-414.

[33] S. A. Goorden, M. Horstmann, A. P. Mosk, B. Škorić, and P. W. H. Pinkse, "Quantum-secure authentication of a physical unclonable key," *Optica,* vol. 1, no. 6, pp. 421-424, 2014/12/20 2014.

[34] S. Sze and K. Ng, "Tunnel Devices," in *Physics of Semiconductor Devices*, 2006, pp. 415-465.

[35] R. Bernardo-Gavito *et al.*, "Extracting random numbers from quantum tunnelling through a single diode," *Scientific Reports,* vol. 7, no. 1, p. 17879, 2017/12/19 2017.

[36] R. M. Iutzi and E. A. Fitzgerald, "Defect and temperature dependence of tunneling in InAs/GaSb heterojunctions," *Applied Physics Letters,* vol. 107, no. 13, p. 133504, 2015.

[37] K. Guan, S. Xie, W. Guo, L. Mao, S. Zhang, and Y. Wang, "Temperature dependence of DC characteristics of resonant tunneling diode," *Guti Dianzixue Yanjiu Yu Jinzhan/Research and Progress of Solid State Electronics,* vol. 33, pp. 428-431+478, 10/01 2013.

[38] O. Vanbésien, R. Bouregba, P. Mounaix, and D. Lippens, "Temperature Dependence of Peak to Valley Current Ratio in Resonant Tunneling Double Barriers," in *Resonant Tunneling in Semiconductors: Physics and Applications*, L. L. Chang, E. E. Mendez, and C. Tejedor, Eds. Boston, MA: Springer US, 1991, pp. 107-116.

[39] M. A. M. Zawawi, K. W. Ian, J. Sexton, and M. Missous, "Fabrication of Submicrometer InGaAs/AlAs Resonant Tunneling Diode Using a Trilayer Soft Reflow Technique With Excellent Scalability," *IEEE Transactions on Electron Devices,* vol. 61, no. 7, pp. 2338-2342, 2014.

[40] M. Hook, "Statistical Techniques and Non-Destructive Testing Methods for Copper Wire Bond Reliability Investigation," 2018.

[41] J. Gomes, "A Study on the Effect of Bond Stress and Process Temperature on Palladium Coated Silver Wire Bonds on Aluminum Metallization," 2015.

[42] TPT, "Wire Bonder HB10/HB16 Operation Manual," ed: TPT, 2016.

[43] I. E. Bagci *et al.*, "Resonant-Tunnelling Diodes as PUF Building Blocks," *IEEE Transactions on Emerging Topics in Computing,* pp. 1-1, 2019.

[44] K.-M. Hwang *et al.*, "Nano-electromechanical Switch Based on a Physical Unclonable Function for Highly Robust and Stable Performance in Harsh Environments," *ACS Nano,* vol. 11, no. 12, pp. 12547-12552, 2017.

[45] S. T. C. Konigsmark, L. K. Hwang, D. Chen, and M. D. F. Wong, "CNPUF: A Carbon Nanotube-based Physically Unclonable Function for secure low-energy hardware design," in *19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2014, pp. 73-78.

[46] B. Astbury *et al.*, "Strong PUFs from arrays of resonant tunnelling diodes," Accessed on: April 01, 2021, Available: https://arxiv.org/abs/1805.03246

[47] T. K. Moon, *Error Correction Coding: Mathematical Methods and Algorithms*. Wiley-Interscience, 2005.

[48] R. Johannesson and K. S. Zigangirov, *Fundamentals of Convolutional Coding*. Wiley-IEEE Press, 2015.

[49] D. Florencio and C. Herley, "A large-scale study of web password habits," presented at the Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, 2007. Available: https://doi.org/10.1145/1242572.1242661

[50] J. Schulman, H. De Los Santos, and D. Chow, "Physics-based RTD current-voltage equation," *Electron Device Letters, IEEE,* vol. 17, pp. 220-222, 06/01 1996.

[51] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition*, 1994, vol. 1, pp. 582-585 vol.1.

[52] D. B. Owen, "A table of normal integrals," *Communications in Statistics - Simulation and Computation,* vol. 9, no. 4, pp. 389-419, 1980/01/01 1980.

Chapter 9: Auxiliary Information

9.1. Appendix 1: Procedural Generation

As an appendix to section 4, it would be valuable to discuss the methodology for the procedural generation of the simulated devices based on the findings therein. To simulate the effectiveness of various binning techniques in section 5, it is necessary to expand out the 256 RTDs (and 12800 measurements) into more continuous distributions from which RTD measurements can be generated and binned. For this, the intra- and inter-distributions are treated independently, such that one can simulate the measurement of RTD by taking a random draw from an inter-measurement distribution (figure part (a)s in the evaluation technique figures in section 4), and then modifying it by a factor based on a random draw from the corresponding normalised intra-measurement distribution (figure part (b)s).

This procedural value generation was performed through two techniques, from two types of distribution. The first and most simple of these was the technique of generating data based on a normal distribution, simply using the means and standard deviations of the intra- and inter-measurement distributions (as also used in the figures of merit ratios), with the inter-measurement mean being based on the mean of the 5-micron devices, and the intra-measurement mean being the value returned by the inter-measurement RTD measurement generation process. This is a simple and efficient technique and can help to directly verify the validity of expressions derived in this work for relating the inter-/intra- ratio of section 4 to the error rate and bias metrics in section 5 via normal approximation. Most binning techniques in the next section are independent of mean values, but for simulation this was taken to be 0.148V and 2.60mA – the mean values for 5 square micron RTD peak voltage and current through the maximum method, respectively.

The second technique for procedurally generating RTD measurements is basing their creation on a non-parametric kernel density estimation (KDE). The distributions used for this (in an often-truncated form) can be seen on the margins of the evaluation technique distributional figures in section 4. This involves replacing each discrete data point with a continuous (here normal) sub-distribution and taking the overall distribution as the superposition of these. This is a more complex technique, and less efficient to simulate, but as this does not negate asymmetries or outlier points it can give a more realistic depiction of the data as taken - and as such a more accurate simulation of the effects of binning on these devices. The KDE can be used to verify the validity in employing a parametric approximation of the real-world RTD measurement distribution, verifying whether the RTD measurement data can be treated analytically via the normal approximation above. For this method the KDE, using a normal kernel function with bandwidth 1, was taken as a cumulative distribution function (CDF) at a resolution of 256,000 points (with linear interpolation between), and was based around the same $5\ \mu\mathrm{m}^2$ mean values as with the generation from a normal distribution.

9.2. Appendix 2: Binarisation Derivations

9.2.1. Single Thresholding Derivations

The following section aims to derive the reliability of the single threshold binning method found in section 5.1. Based on symmetry, and taking the threshold location to be at the mean of the inter-measurement distribution, the reliability can be described as:

$$R = 2\int_{\mu_i=-\infty}^{0} f_o(\mu = 0, \sigma = \sigma_o; x = \mu_i)\left[\int_{x_i=-\infty}^{0} f_i(\mu = \mu_i, \sigma = \sigma_i; x = x_i)\, dx_i\right] d\mu_i$$

Where:

$$\left[\int_{x_i=-\infty}^{0} f_i(\mu = \mu_i, \sigma = \sigma_i; x = x_i)\, dx_i\right] = F_i(0) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{0-\mu_i}{\sigma_i\sqrt{2}}\right)\right]$$

And where f and $\operatorname{erf}(\mathrm{x})$ represent the normal probability density function (PDF) and error function of x respectively. Here the subscript *i* represents association with the intra-measurement distribution (as in inner) and subscript *o* represents an association with the inter-measurement distribution (as in outer). μ, σ and x represent the mean, standard deviation, and PDF location variable, respectively. These can be converted into the standard Gaussian PDF and CDF functions, $\varphi(\mathrm{x})$ and $\phi(\mathrm{x})$ respectively:

$$f_o(x = \mu_i | \mu_o = 0, \sigma_o) = \frac{1}{\sigma_o}\varphi_o\left(\frac{\mu_i}{\sigma_o}\right)$$

$$F_i(x = 0 | \mu_i, \sigma_o) = \phi\left(\frac{0-\mu_i}{\sigma_i}\right) = \phi\left(\frac{-\mu_i}{\sigma_i}\right)$$

Leading to an integral:

$$R = \frac{2}{\sigma_o}\int_{\mu_i=-\infty}^{0} \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{-\mu_i}{\sigma_i}\right) d\mu_i = \frac{2}{\sigma_o}\int_{\mu_i=-\infty}^{0} \varphi_o(a\mu_i)\,\phi(b\mu_i)\, d\mu_i$$

Where $a = 1/\sigma_o$ and $b = -1/\sigma_i$, which is in the form of the identity:

$$\int_{-\infty}^{0} \varphi_o(ax)\,\phi(bx)\, dx = \frac{1}{2\pi|a|}\left(\frac{\pi}{2} - arctan\left(\frac{b}{|a|}\right)\right)$$

Given $\tan^{-1}(-x) = -\tan^{-1}(x)$ and cancelling where appropriate, we get the result:

$$R = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{\sigma_i}\right)$$

A second, more generic analysis can be done where the threshold is allowed to be set arbitrarily. This would not ensure symmetry, and so bias will vary, as will the error rate in a more complex fashion. In this situation, taking T as the now non-zero threshold finds the bias as:

$$\beta_0 = F(T) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{T}{\sigma_o\sqrt{2}}\right)\right)$$

Where β_0 is the bias of the leftmost side, here called bit 0. $\beta_0 = 1 - \ \beta_1$, where β_1 stands for the bias of the rightmost side.

The error rate for this generic solution introduces the cumulative bivariate normal distribution BvN. First, the reliability for the leftmost bit state (here denoted 0) can be written as:

$$R_0 = \int_{\mu_i=-\infty}^{T} f_o(\mu = 0, \sigma = \sigma_o; x = \mu_i)\left[\int_{x_i=-\infty}^{T} f_i(\mu = \mu_i, \sigma = \sigma_i; x = x_i)\, dx_i\right] d\mu_i$$

or

$$R_0 = \int_{\mu_i=-\infty}^{T} f_o(\mu_i)\, F_i(T)\, d\mu_i = \int_{\mu_i=-\infty}^{T} f_o(\mu_i)\, F_i(T)\, d\mu_i = \frac{1}{\sigma_o}\int_{\mu_i=-\infty}^{T} \varphi_o\left(\frac{\mu_i}{\sigma_o}\right) \phi\left(\frac{T-\mu_i}{\sigma_i}\right) d\mu_i$$

Which, changing variables to $x = \ \mu_i/\sigma_o$, gives:

$$R_0 = \int_{x=-\infty}^{T/\sigma_o} \varphi_o(x)\, \phi\left(\frac{T-\sigma_o x}{\sigma_i}\right) dx$$

Which is in the form of the identity [52]:

$$\int_{x=-\infty}^{Y} \varphi_o(x)\, \phi(a+bx)\, dx = BvN\left[\frac{a}{\sqrt{1+b^2}}, Y; \rho = \frac{-b}{\sqrt{1+b^2}}\right]$$

Where $a = T/\sigma_i$, $b = -\sigma_o/\sigma_i$ and $Y = T/\sigma_o$. As T is also scaled from any absolute value by the transform here affecting the normal distributions, a and Y are ratios comparable to σ_o/σ_i ratios. This leads to a reliability of:

$$R_0(T) = BvN\left[\frac{T}{\sigma_o\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}}, \frac{T}{\sigma_o}; \rho = \frac{\sigma_o}{\sigma_i\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}}\right]$$

Where:

$$BvN(h,k;\rho) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\int_{-\infty}^{k}\int_{-\infty}^{h} exp\left(-\left(\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right)\right)dx\,dy$$

And due to the symmetry of the two distributions around 0, we can find the reliability of the rightmost half (here denoted 1) as equal to the leftmost half function, with the relationship $T_o = -T_1$:

$$R_1(T) = R_0(-T) = BvN\left[\frac{-T}{\sigma_o\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}},\frac{-T}{\sigma_o};\rho = \frac{\sigma_o}{\sigma_i\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}}\right]$$

And the overall reliability of this mechanism is $R = R_0 + R_1$:

$$R = BvN\left[\frac{T}{\sigma_o\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}},\frac{T}{\sigma_o};\rho = \frac{\sigma_o}{\sigma_i\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}}\right] + BvN\left[\frac{-T}{\sigma_o\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}},\frac{-T}{\sigma_o};\rho = \frac{\sigma_o}{\sigma_i\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}}\right]$$

With fewer symmetries to exploit, this equation is far less approachable than for the case with the threshold at the mean, but as a non-optimal implementation case it is here for completeness only.

9.2.2. Pairwise Comparison Derivations

This section aims to derive the reliability for the pairwise comparison binning technique introduced in section 5.2. The reliability of the technique can be expressed as:

$$R = 2\cdot P(N_i < N_0 \cap N_i < 0)$$

Where $N_0 \sim N(0{,}2\sigma_o^2)$ represents the difference in the real, inter-measurement positions of comparative elements 1 and 2 and $N_i \sim N(0{,}2\sigma_i^2)$ the difference from measurement errors of the measurement of the two elements.

$$R = 2\int_{\mu_i=-\infty}^{0} f_{N_o}\big(\mu = 0,\sigma = \sqrt{2}\sigma_o; x = \mu_i\big)\left[\int_{x_i=-\infty}^{-\mu_i} f_{N_i}\big(\mu = 0,\sigma = \sqrt{2}\sigma_i; x = x_i\big)\,dx_i\right]d\mu_i$$

Where:

$$\left[\int_{x_i=-\infty}^{-\mu_i} f_i(\mu = \mu_i,\sigma = 2\sigma_i; x = x_i)\,dx_i\right] = F_i(-\mu_i) = \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{-\mu_i}{2\sigma_i}\right)\right]$$

$$f_o(x = \mu_i|\mu_o = 0,\sigma_o) = \frac{1}{\sqrt{2}\sigma_o}\varphi_o\left(\frac{\mu_i}{\sqrt{2}\sigma_o}\right)$$

$$F_i(x = 0|\mu_i, \sigma_o) = \phi\left(\frac{0 - \mu_i}{\sqrt{2}\sigma_i}\right) = \phi\left(\frac{-\mu_i}{\sqrt{2}\sigma_i}\right)$$

Leading to an integral:

$$R = \frac{2}{\sqrt{2}\sigma_o}\int_{\mu_i=-\infty}^{0} \varphi_o\left(\frac{\mu_i}{\sqrt{2}\sigma_o}\right)\phi\left(\frac{-\mu_i}{\sqrt{2}\sigma_i}\right)\,d\mu_i = \frac{2}{\sigma_o}\int_{\mu_i=-\infty}^{0} \varphi_o(a\mu_i)\,\phi(b\mu_i)\,d\mu_i$$

Where $a = 1/\sqrt{2}\sigma_o$ and $b = -1/\sqrt{2}\sigma_i$, which is in the form of the identity:

$$\int_{-\infty}^{0} \varphi_o(ax)\,\phi(bx)\,dx = \frac{1}{2\pi|a|}\left(\frac{\pi}{2} - arctan\left(\frac{b}{|a|}\right)\right)$$

Given $\tan^{-1}(-x) = -\tan^{-1}(x)$ and cancelling where appropriate, we get the result:

$$R = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_o}{\sigma_i}\right)$$

9.2.3. Multiple Thresholding Derivations

The final derivation in this appendix aims to derive the bias and reliability in the much more complicated case of multiple threshold binning, as in section 5.3. When the centremost threshold is exactly at the mean $\mu_o = 0$, both sides are symmetrically alternating in bit state – in other words for any given interval on the positive side past the mean, there is an equal interval of the opposite bit state on the other side of the mean, resulting in perfectly equal biasing. Taking the worst-case offset, where a single bit state is centrally located at the mean (in other words the halfway point of an interval is at the mean), we can best calculate the bias using iterative confidence intervals.

The interval of probability ($\tau(x)$) of an element being between $\pm x$ in a centred ($\mu = 0$) normal distribution is:

$$\tau(x) = \mathrm{F}(x) - \mathrm{F}(-x) = \phi\left(\frac{x}{\sigma}\right) - \phi\left(-\frac{x}{\sigma}\right) = \mathrm{erf}\left(\frac{x}{\sigma\sqrt{2}}\right)$$

And the difference in probabilities between the two states $\Delta\beta$ is:

$$\Delta\beta = \sum_{n=1}^{\infty} \tau(2nI) - \tau((2n-1)I) = \sum_{n=1}^{\infty} \mathrm{erf}\left(\frac{2nI}{\sigma_o\sqrt{2}}\right) - erf\left(\frac{(2n-1)I}{\sigma_o\sqrt{2}}\right)$$

Where the probability of the bit associated with the central probability interval (let's say β_A) is:

$$\beta_A = \frac{1}{2}(1 + \Delta\beta)$$

The error rate for this technique combines the intra- and inter-measurements in a way that adds further complexity. First, the overall probability weight of intra-measurement distributions with means inside segments A and B is:

$$R_{AB} = \int_B^A f_o(\mu_i)\big(F_i(B) - F_i(A)\big)\, d\mu_i$$

Then, using a similar process as with the single threshold case:

$$R_{AB} = \frac{1}{\sigma_o}\int_B^A \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{B-\mu_i}{\sigma_i}\right)\, d\mu_i - \frac{1}{\sigma_o}\int_B^A \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{A-\mu_i}{\sigma_i}\right)\, d\mu_i$$

This can be expanded out with integral limits to become four terms:

$$\begin{aligned} R_{AB} = {} & \frac{1}{\sigma_o}\int_{-\infty}^A \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{B-\mu_i}{\sigma_i}\right)\, d\mu_i - \frac{1}{\sigma_o}\int_{-\infty}^B \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{B-\mu_i}{\sigma_i}\right)\, d\mu_i \\ & - \frac{1}{\sigma_o}\int_{-\infty}^A \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{A-\mu_i}{\sigma_i}\right)\, d\mu_i + \frac{1}{\sigma_o}\int_{-\infty}^B \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{A-\mu_i}{\sigma_i}\right)\, d\mu_i \end{aligned}$$

Each of which is in the form of the cumulative bivariate normal expressed prior, with differing thresholds and CDF positions. For ease of notation, we will express the cumulative bivariate normal as the following modified form:

$$BvN(L,M) = \frac{1}{\sigma_o}\int_{-\infty}^M \varphi_o\left(\frac{\mu_i}{\sigma_o}\right)\phi\left(\frac{L-\mu_i}{\sigma_i}\right)\, d\mu_i = BvN\left[\frac{L}{\sigma_o\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}}, \frac{M}{\sigma_o}; \rho = \frac{\sigma_o}{\sigma_i\sqrt{1+\left(\frac{\sigma_o}{\sigma_i}\right)^2}}\right]$$

Applying this to R_{AB} gives:

$$R_{AB} = BvN(B,A) + BvN(A,A) - BvN(B,B) - BvN(A,B)$$

Where A and B are the start and end points of a given interval on the inter distribution. To give these points meaning, we will set $A_0 = (n+o)I$ and $B_0 = (n+o+0.5)I$, where the offset o is in the domain $0 \leq o < 0.5$ and n is the index of a sum from negative to positive infinity. This will give the reliability metric for a single parity of bit, here called 0. That is to say:

$$\begin{aligned} R_0(o,I) &= BvN(B_0,A_0) + BvN(A_0,A_0) - BvN(B_0,B_0) - BvN(A_0,B_0) \\ &= BvN((n+o+0.5)I,(n+o)I) + BvN((n+o)I,(n+o)I) \\ &\quad - BvN((n+o+0.5)I,(n+o+0.5)I) - BvN((n+o)I,(n+o+0.5)I) \end{aligned}$$

And where the other parity of bit R_1 is given by the other set of intervals, where $A_1 = (n+o+0.5)I$ and $B_1 = (n+o+1)I$. It is worth noting here that B_0 and A_1 are equal, which slightly simplifies the total equation for reliability $R = R_0 + R_1$:

$$R = \sum_{n=-\infty}^{\infty} BvN(B_0, A_0) + BvN(A_0, A_0) - BvN(A_0, B_0) + BvN(B_1, A_1) - BvN(B_1, B_1) - BvN(A_1, B_1)$$

The reliability of the method can therefore be written as:

$$R = \sum_{n=-\infty}^{\infty} BvN(B_0, A_0) + BvN(A_0, A_0) - BvN(A_0, B_0) + BvN(B_1, A_1) - BvN(B_1, B_1) - BvN(A_1, B_1)$$

Where set $A_0 = (n + o)I$, $B_0 = (n + o + 0.5)$, $A_1 = (n + o + 0.5)I$ and $B_1 = (n + o + 1)I$, with I representing the thresholding interval as before, and o representing a variable offset of the threshold lines with respect to the inter-measurement distribution in the domain $0 \leq o < 0.5$. BvN represents the bivariate normal, where:

$$BvN(L, M) = BvN\left[\frac{L}{\sigma_o\sqrt{1 + \left(\frac{\sigma_o}{\sigma_i}\right)^2}}, \frac{M}{\sigma_o}; \rho = \frac{\sigma_o}{\sigma_i\sqrt{1 + \left(\frac{\sigma_o}{\sigma_i}\right)^2}}\right]$$

This converges, but cannot be written any more simply, and would require numerical techniques to evaluate. The offset value o should be chosen as the worst case for the reliability of a given σ_i and σ_o. Similarly, the infinite series can be reduced down to finite either by taking the limits of the measurement window or by limiting the series to include, for instance, 99% of the total probability as a cut-off.

9.3. List of Acronyms

ADC	Analogue to digital converter
CDF	Cumulative distribution function
CRP	Challenge-response pair
DAC	Digital to analogue converter
FNR	False negative rate
FPR	False positive rate
IV	Current-Voltage (domain, as for electronic characterisation)
KDE	Kernel density estimation
MA	Moving average (smoothing algorithm)
MBE	Microwave beam epitaxy
NDR	Negative differential resistance (region of IV relationship)
PDF	Probability density function
PDR	Positive differential resistance (region of IV relationship)
PUF	Physical unclonable function
RTD	Resonant tunnelling diode