Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

Domain Generalization for Depolarizing Chipless-RFID Tag Identification under Position Shift

Access-aware evaluation and failure localization under an unseen acquisition geometry

Dissertation presented by

Yangdeyi Yang

Student ID: 124134230

for the degree of

Master of Engineering Science (MEngSc)
in Electrical and Electronic Engineering

University College Cork

School of Engineering and Architecture
Department of Electrical and Electronic Engineering

Head of School: **Dr Alan P. Morrison**

Supervisors
Dr Salvatore Tedesco
Prof. Liam Marnane

Research conducted in collaboration with Tyndall National Institute
EE6023 — Dissertation in Electrical and Electronic Engineering
Dissertation submitted in part fulfilment of the requirements of the MEngSc in Electrical and Electronic Engineering

September 2026

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Signed: Yangdeyi Yang Date: 8 September 2026

Acknowledgements

I would like to express my sincere gratitude to Dr Salvatore Tedesco for his patient and continuous supervision throughout this project. His guidance was central to shaping the direction of the research, keeping the work focused on the most important scientific questions, and refining the experimental programme as it developed. I am particularly grateful for his detailed comments on successive drafts of the dissertation, his careful review of the experimental code and implementation, and for making the original measurement datasets available and helping me understand their experimental context and provenance. His advice, from the overall research direction to the technical details of individual analyses, was invaluable in bringing this dissertation to its final form.

I would also like to thank Prof. Liam Marnane for his continued support, perspective, and constructive advice throughout the project. His careful reading of the dissertation and his comments on its structure, technical presentation, and clarity were extremely valuable in strengthening the final manuscript. I am also grateful to Tyndall National Institute for providing the research environment and resources that supported this work.

I would also like to express my deepest gratitude to my parents for their unwavering support, encouragement, and understanding throughout my studies. Their belief in me and the support they have given me over the years have provided a constant source of motivation, not only during this project but throughout my academic journey. I am profoundly grateful for everything they have done to make this journey possible.

Finally, I would like to thank my girlfriend, Amethy, for her companionship, patience, and constant support throughout this demanding period. Her encouragement through the long stretches of experimentation, coding, debugging, and writing made the process far more manageable, and I am deeply grateful to have had her beside me throughout it.

Abstract

Chipless radio-frequency identification replaces a silicon identifier with a passive resonant structure. That lowers tag complexity, but it makes identity inseparable from the measurement: reader angle, distance, material loading and noise all reshape the spectrum presented to the classifier. Earlier work showed that the tags were learnable when training and test measurements came from the same varied campaign. This dissertation asks the next engineering question: does that recognition survive at a reader geometry excluded from model development?

The principal corpus contains 12,600 cross-polar spectra from seven tag identities measured at three permittivity levels, on three surfaces and at four reader geometries. P1 is 50 mm at 0°, P2 is 50 mm at 45°, P3 is 150 mm at 0°, and P4 is 150 mm at 45°. Training uses P1–P3 and evaluation uses P4, the unobserved joint long-distance/45° corner. Fifty repeated sweeps of each physical condition are kept together, so the analysis is based on 252 condition blocks rather than treating 12,600 rows as independent experiments.

A first-difference convolutional model selected using only P1–P3 reached P4 Accuracy 0.1566 and Macro-F1 0.1122 across five seeds, close to the balanced uniform-random reference of 1/7, and every seed abandoned at least one class. Three standard source-only remedies—invariant risk minimisation, domain-adversarial training and group distributionally robust optimisation—did not reliably improve their matched controls. A proposed cross-position Mixup study could not satisfy its pairing rule and therefore produced no performance result. Limited target calibration, including few-shot support, a trainable readout and encoder fine-tuning, gave only small or unstable recovery.

The diagnostics place the failure at three connected levels. Performance degrades at the two 45° positions before learned encoding; at P4 the learned representation preserves less linearly accessible TagID discrimination than the raw spectrum; and changing the readout recovers only a small, uncertain amount. The evidence does not isolate one causal mechanism, but it shows why a model that succeeds on familiar measurements can fail at a new installation. The main engineering implication is that robust chipless-RFID identification requires measurement design, representation design and evaluation design to be developed together.

Keywords: chipless RFID; depolarizing tag; domain generalisation; position shift; acquisition-factor analysis; representation diagnostics; few-shot adaptation; evaluation protocol; 1-D convolutional network; nearest-class-mean.

Table of Contents

List of Figures

List of Tables

Abbreviations and Notation

Term	Meaning
Block	a condition block: one tag identity at one permittivity class on one surface at one position; 50 repeated sweeps
CNN	convolutional neural network
CORAL	correlation alignment; a second-order feature-alignment objective
DANN	domain-adversarial neural network; uses a gradient-reversal layer
DG	domain generalisation: training on source domains and testing on an unseen one
ER	relative-permittivity class of the material beneath the tag (0, 1, 2)
ERM	empirical risk minimisation
FDTD	finite-difference time-domain electromagnetic simulation
GRL	gradient-reversal layer
GroupDRO	group distributionally robust optimisation
IRM	invariant risk minimisation; IRM is the practical penalty-based form
LOPO	leave-one-position-out
Macro-F1	unweighted mean of the per-class F1 scores; here the primary metric because it is sensitive to class collapse
NCM	nearest-class-mean readout using class prototypes
P1–P4	reader positions: P1 = 50 mm / 0°, P2 = 50 mm / 45°, P3 = 150 mm / 0°, P4 = 150 mm / 45°
RCS	radar cross section
Retrospective target-informed	full target outcomes influenced a decision; never Strict-DG
SD	standard deviation; population (ddof = 0) unless stated. Seed SD is dispersion, not a confidence interval
SSL	self-supervised learning
Strict source-only	an evaluation in which no target-position information influenced any fitting, selection or hyper-parameter decision
Target-assisted	labelled target support or validation data are used under an explicit protocol
Target-informed	an evaluation in which target-position outcomes influenced at least one decision; not domain generalisation
VNA	vector network analyser
Within-condition interpolation	target partitions share physical conditions or near-duplicates; not held-condition transfer
Zero-recall class	a class that receives no correct prediction at all

1 Introduction

1.1 Motivation

A conventional radio-frequency identification (RFID) tag contains a silicon integrated circuit. That circuit stores the identifier and modulates the backscattered field. Chipless RFID removes it, which is the central economic motivation for the technology. Identity is instead encoded in the geometry of a purely passive structure, so that the tag becomes a printed or etched pattern whose electromagnetic response is its identifier [1], [2]. The economic argument is straightforward: a printed or etched tag may be viable for high-volume items for which a chipped tag is uneconomic.

That saving shifts complexity to the reader. In a chipped system the tag actively participates in the protocol: it decodes a query, retrieves a stored word and replies. The reader therefore receives a digital message. In a chipless system the tag does nothing except scatter. The reader receives an analogue electromagnetic signature and must infer identity from it. Everything that lies between the transmitting antenna and the receiving antenna — the propagation path, the material beneath the tag, the curvature of the mounting surface, the orientation and distance of the reader, and scattering from nearby objects — is superimposed on that signature. The intelligence that a chip would have supplied must therefore be relocated to the reader, and it must remain effective when those conditions change.

This relocation makes chipless-RFID identification a machine-learning problem, but it also creates a fundamental ambiguity. A classifier trained on signatures collected in one measurement configuration may be learning tag identity, or it may partly be learning the configuration in which that identity was observed. As long as training and testing draw from familiar configurations, the two can be difficult to distinguish from the accuracy figure alone. A model that performs well in the laboratory may therefore still fail when the reader-to-tag geometry changes at deployment. The central engineering issue is not only whether the tags are learnable, but whether the learned decision remains valid when the measurement configuration is no longer familiar. Figure 1 maps the resulting scientific argument.

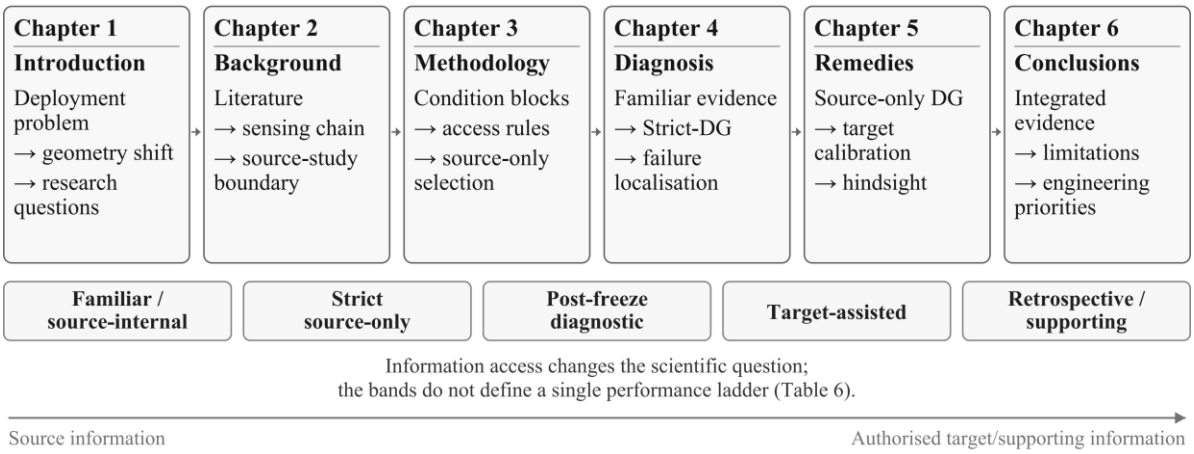

Figure 1. Reader-oriented map of the dissertation, showing the progression from the physical sensing problem to strict transfer, failure diagnosis, controlled remedies and engineering conclusions.

1.2 The engineering problem: position shift as domain shift

Reader geometry is represented in the principal measurement campaign by four discrete reader positions, each defined by two physically meaningful acquisition factors: reader-to-tag distance and acquisition angle. Changing either factor can alter the electromagnetic signature presented to the classifier. A deployment condition in which both factors take a combination absent from model development is therefore different from another random measurement drawn from an already represented configuration.

The experiment provides a compact 2 × 2 geometry design. The four positions are P1 = 50 mm at 0°, P2 = 50 mm at 45°, P3 = 150 mm at 0°, and P4 = 150 mm at 45°. P1–P3 provide the development data, while P4 is reserved as the held target geometry. The source positions expose the learner to both levels of distance and both levels of angle individually, but never to their joint long-distance/45° setting. P4 is therefore the held joint 150-mm/45° geometry in the governed Strict-DG evaluation. Its role is defined by this information boundary rather than by an assumption that it is intrinsically the most difficult of the four positions.

In machine-learning terminology, each reader position is treated as a domain. The physical geometry changes the distribution of the measured input while the TagID label space remains unchanged. Training and selecting a model using P1, P2 and P3, then evaluating it at P4 without allowing labelled or unlabelled P4 data to influence development, defines the primary **strict source-only domain-generalisation (Strict-DG)** problem [5], [21]. It represents the practical situation in which a reader and recognition model are characterised under a limited set of laboratory geometries and are subsequently required to operate at a geometry that did not participate in model development.

Two aspects make this particular Strict-DG problem especially demanding. First, there are only three source domains, so methods that attempt to infer stable or invariant structure across environments have relatively little environmental variation from which to estimate it. Second, transfer to P4 requires generalisation to the held joint factor combination rather than interpolation among observed geometries. The question is consequently not simply whether the classifier can recognise new measurement rows, but whether TagID information learned from the three source geometries remains usable at a held target geometry.

Figure 2 provides a physical reference for the passive depolarizing tag studied throughout this dissertation; the detailed resonant mechanism is introduced in Section 2.2.

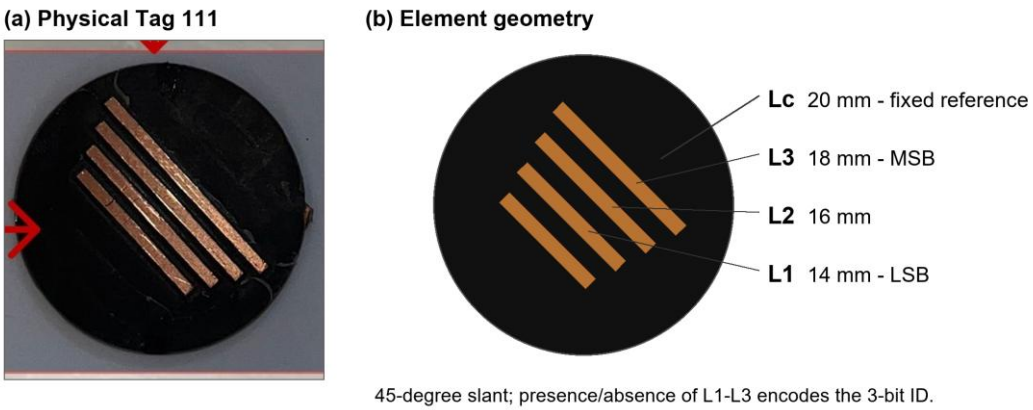

Figure 2. Physical Tag 111 from the principal depolarizing-tag family, shown with its resonant-element geometry. Adapted from [4].

1.3 Why the evaluation protocol is the central difficulty

A recurring theme of this dissertation is that, on repeated-measurement sensing data, the evaluation protocol determines what a performance score can legitimately mean. Two mechanisms are particularly important.

The first is repeated measurement. Each physical condition is measured fifty times, so the 12,600 rows contain many highly related repeated spectra. If individual rows are randomly divided between training and test sets, measurements from the same physical condition can appear on both sides of the split. The resulting score therefore primarily reflects interpolation among familiar physical conditions rather than generalisation to previously unseen ones. For this reason, the principal held-condition and transfer evaluations keep all repeated measurements belonging to the same TagID–permittivity–surface–position condition together whenever a split is constructed. Experiments that deliberately permit shared physical conditions are identified separately as within-condition diagnostics. The formal definition of these condition blocks is given in Section 3.1.

The second mechanism is model selection. A Strict-DG claim has no target-domain validation set: once target information influences the choice of representation, architecture, hyper-parameters or checkpoint, the resulting experiment answers a different, target-informed question. Gulrajani and Lopez-Paz showed that conclusions in domain generalisation can depend strongly on the model-selection protocol and that carefully tuned empirical risk minimisation remains a strong baseline under consistent evaluation procedures [6]. In the present dissertation, a model is therefore not promoted because it happens to perform better on P4. Strict-DG models are selected using source evidence, and P4 is consulted only at the stage permitted by the declared evaluation regime.

This distinction becomes important because later experiments intentionally use different levels of information access. Familiar-condition evaluation asks whether TagID is learnable when the relevant measurement conditions are already represented. Strict-DG asks whether a model developed on P1–P3 survives at P4. Post-freeze diagnostics use held data to investigate the structure of an already fixed model without retroactively altering its selection. Target-assisted experiments ask what changes when limited P4 information is deliberately authorised. Retrospective and supporting studies answer still different questions. These results are therefore interpreted within their own information-access regimes rather than combined into a single performance leaderboard.

1.4 Research questions

The dissertation is organised around five questions that build from the evaluation problem to transfer, diagnosis, remedies and engineering implications. Table 1 links each question to its primary location and decision evidence.

Table 1. Research questions, primary locations and evidence used to answer them.

#	Research question	Primary location	Decision evidence
RQ1	Why do familiar-condition results differ from transfer to a held reader geometry?	Section 3.1; Chapter 4	Compare the data grain, grouping rule and matched within-position evidence.
RQ2	Can a recipe developed on P1–P3 generalise to the held P4 geometry, and was its source-only selection defensible?	Section 3.2; Chapter 4	Use the source-only selector, frozen five-seed Accuracy, Macro-F1, class-collapse and source-selection-regret evidence.
RQ3	Where is the transfer failure located across the measured signal, learned representation and readout?	Chapter 4; Sections 5.2 and 6.1.2	Interpret the acquisition-factor, RAW/first-difference/embedding and controlled-readout diagnostics jointly.
RQ4	Can standard DG objectives or controlled target information reliably repair the failure?	Sections 5.1–5.3	Use matched effects for source-only interventions and access-matched target experiments; do not combine the two access classes into one leaderboard.
RQ5	What do the findings imply for future chipless-RFID measurement and system design?	Chapter 6; Appendices E–F	Separate conclusions supported by the principal experiment from external-corpus and simulation evidence.

1.5 Contributions

This dissertation makes four contributions to the evaluation and diagnosis of geometry shift in chipless-RFID recognition.

1. **Protocol formulation and strict transfer evaluation.** It establishes a block-aware P1–P3 → P4 evaluation in which repeated measurements of the same physical condition remain together and P4 is excluded from preprocessing, model fitting and source-only model selection. This provides an interpretable estimate of transfer to a held reader geometry rather than performance within familiar physical conditions (Sections 3.1–3.2; Chapter 4).
2. **Empirical separation of familiar-condition performance from held-geometry transfer.** It shows that strong performance when measurement conditions are represented in development data does not imply successful transfer to the held P4 geometry. A source-side selection analysis additionally provides descriptive support for the final source-selected representation within the recorded candidate/score set and examines the limitations of that selection rule through source-score reuse (Chapter 4).
3. **Three-level localisation of the transfer failure.** Rather than treating the P4 result as a single unexplained accuracy reduction, the dissertation separates geometry-associated changes in the measured signal, loss of linearly accessible TagID discrimination in the learned representation, and residual readout mismatch. The resulting evidence supports a joint physical and representational interpretation without claiming a single causal electromagnetic mechanism (Chapter 4; Section 5.2).

4. **Controlled evaluation of remedy and calibration boundaries.** It compares standard source-only domain-generalisation interventions with progressively more informative target-access experiments while keeping their information boundaries explicit. The experiments establish which tested interventions did not provide a reliable rescue under the present protocols and what limited target calibration can — and cannot — recover (Chapter 5).

Together, these contributions shift the emphasis from reporting a single best classification score to determining what kind of generalisation is actually being measured, where transfer fails, and what additional information is required before that failure can be repaired.

1.6 Scope and boundaries of the claims

The evidence in this dissertation supports deliberately bounded conclusions. These boundaries are stated explicitly so that diagnostic, target-assisted or exploratory results are not interpreted as stronger generalisation evidence than their protocols allow.

1. **Domain generalisation.** The experiments establish the empirical limitations of the selected empirical-risk model and the tested IRM, DANN and GroupDRO interventions under the present three-source, held-P4 protocols. They do not establish that domain generalisation is intrinsically impossible for chipless-RFID recognition. The pre-specified Mixup study contributes feasibility evidence only because its pairing rule could not be satisfied without changing the frozen protocol.
2. **Physical mechanism.** The acquisition-factor and spectral analyses identify associations between reader geometry and recognition behaviour. They support physically plausible interpretations of the failure, but they do not isolate a causal electromagnetic mechanism.
3. **Target-informed evidence.** Target-assisted and retrospective analyses are used to study recoverability once additional P4 information is permitted. Their results are not Strict-DG estimates and are not compared with the strict P4 result as though they answered the same question.
4. **Physical redesign.** The electromagnetic simulation study is exploratory. Its confirmation criteria were not met, so the dissertation does not claim a validated tag redesign or make a fabrication recommendation from that branch.
5. **State-of-the-art comparison.** No shared external benchmark currently provides a like-for-like chipless-RFID position-shift task with the same tag identities,

measurement factors and target geometry. The dissertation therefore does not present its results as a state-of-the-art leaderboard comparison.

These limitations define the scope of the claims rather than weakening them: conclusions are made only at the level supported by the corresponding information access and experimental evidence.

1.7 Structure of the dissertation

The remainder of the dissertation follows the sequence established by the research questions. Chapter 2 explains the relevant literature, the physical sensing system and the point of departure from the source study. Chapter 3 translates that physical model into the dataset structure used by the experiments, defines condition blocks, information-access boundaries and the source-only model-selection rule, and establishes the metrics, inferential units, uncertainty procedures and reproducibility controls required to interpret the later results. Chapter 4 establishes the difference between familiar-condition performance and held-geometry transfer and then diagnoses the P1–P3 → P4 failure through source-selection, acquisition-factor and representation analyses. Chapter 5 tests possible remedies, progressing from source-only domain-generalisation objectives to controlled target calibration and retrospective hindsight, followed by an information-access synthesis. Chapter 6 integrates the evidence, states the remaining limitations and draws the resulting engineering conclusions and priorities for future work. External-dataset studies, simulation evidence and detailed reproducibility material are retained in the appendices so that they support, rather than obscure, the main argument.

2 Background and Point of Departure

This chapter establishes the background required to understand the experimental programme. Section 2.1 positions the work within the chipless-RFID and machine-learning literature. Section 2.2 then follows the physical sensing process from the resonant tag to the spectral representation presented to the learning system. Section 2.3 closes the chapter by distinguishing what the source study had already established from the held-geometry question addressed in this dissertation.

2.1 Technical Background and Related Work

This section positions the dissertation within four connected strands of prior work. It first reviews how frequency-coded and depolarizing chipless-RFID systems encode and recover identity, then considers the progression from signal-processing methods to machine-learning decoders and the evaluation protocols used to assess them. It next introduces domain generalisation, model selection and target adaptation as the machine-learning framework for geometry shift. The section concludes by identifying the specific gap addressed in this dissertation. Detailed explanation of the physical sensing chain itself is deferred to Section 2.2.

2.1.1 Chipless-RFID sensing and robustness

Frequency-domain chipless RFID is often described as an electromagnetic barcode. Instead of storing an identifier in silicon, the tag uses resonant structures whose geometry produces characteristic features in the measured frequency response. Individual resonators may contribute amplitude peaks, notches or phase changes at designed frequencies, so that combinations of resonant elements form a spectral code [1], [7]. Preradovic et al. demonstrated a 35-bit multiresonator tag in which spiral resonators contributed amplitude and phase features [7]. Subsequent designs increased coding density through combinations of frequency, amplitude and phase states and through fully printable multiresonant structures [8], [9]. Reviews of the field consistently identify reliable detection outside controlled measurement conditions as a continuing challenge [2], [10], [11].

One response to environmental scattering has been the development of depolarizing tags and cross-polar interrogation. A depolarizing structure redirects part of the incident energy into the orthogonal polarisation, allowing the receiver to suppress a substantial component of the co-polar background response [12], [13]. Related dual-polarised approaches have also reduced the need for conventional background calibration [14]. These developments improve the observability of the tag response, but they do not make that response

independent of acquisition geometry. The measured cross-polar signature can still change with reader orientation, distance, material loading and the interaction between the tag and its surroundings. Robust identification therefore remains partly a problem of recognising the same physical tag through measurements that are not invariant across deployment conditions.

This distinction is important for the present dissertation. The relevant question is not whether depolarizing tags provide a cleaner measurement than conventional co-polar readout; that has already been established in the literature. The unresolved question is whether the identity-bearing structure that remains after cross-polar measurement is sufficiently stable for a learned decoder to transfer to a reader geometry that was absent during model development. Section 2.2 develops the physical basis of that question in more detail.

2.1.2 Machine-learning decoding and evaluation under measurement shift

Robust chipless-RFID decoding has been approached through both signal processing and machine learning. Earlier signal-processing methods include dual-polarised, normalisation-free interrogation [14] and continuous-wavelet-transform peak extraction [15]. Machine-learning studies subsequently moved from classifiers operating on engineered features toward learned representations and convolutional networks. Sokoudjou et al. developed measurement-to-identification pipelines using classical learning methods and later examined algorithm choices for software-defined-radio readers and time–frequency representations [16]–[18]. On the Tyndall datasets, deep-learning identification was demonstrated first for a sensor-tag system and subsequently for the depolarizing-tag measurement campaign studied here [3], [4], with later work considering more compact one-dimensional convolutional architectures [19].

These studies establish an important point: chipless-RFID spectra contain substantial information about tag identity when training and evaluation measurements are drawn from the same acquisition population. What they do not automatically establish is transfer to a physical condition that was excluded from model development. This distinction matters because experimental RF datasets commonly contain repeated measurements of the same static physical condition. If individual measurement rows are divided between training and test sets without grouping by physical condition, repeated measurements belonging to the same condition can appear on both sides of the split. Such an evaluation primarily tests interpolation among familiar physical conditions rather than recognition under a genuinely held condition.

The distinction between within-dataset evaluation and broader generalisation is also visible in recent chipless-RFID work. Sokoudjou et al. reported accuracy differences of up to 0.23 between evaluation on a held-out subset of a development dataset and evaluation on a genuinely distinct dataset [20]. That result does not identify the same physical mechanism studied in this dissertation, because cross-dataset changes may involve many factors simultaneously. It nevertheless demonstrates the broader concern that strong performance within a familiar measurement population need not survive a substantial change in the acquisition distribution.

The present work examines that concern in a more controlled setting. Rather than changing the complete dataset, it asks what happens when an entire reader geometry is withheld while the tag identities and principal measurement campaign remain fixed. Repeated measurements of one physical condition are kept together so that the resulting test asks about transfer across conditions rather than interpolation between repeated observations of conditions already represented during development.

2.1.3 Domain generalisation, model selection and target adaptation

When labelled data are available from several source environments but the deployment environment is unavailable during model development, the problem is commonly formulated as domain generalisation [5], [21]. A broad range of methods has been proposed, but they pursue robustness in different ways. CORAL reduces differences in second-order feature statistics across domains [22], while discrepancy-based approaches compare broader feature distributions [23]. Domain-adversarial learning such as DANN encourages representations from which the source domain is difficult to predict [24]. Invariant risk minimisation instead seeks predictors whose optimality is stable across environments [25], while GroupDRO emphasises performance on the worst-performing training group [26]. Input-space interpolation methods such as Mixup provide another route by generating intermediate training examples rather than explicitly aligning learned feature distributions [27].

These approaches should not be treated as implementations of one identical invariance assumption. They differ substantially in objective and mechanism. They nevertheless face a common difficulty in physical sensing: changes in the environment may affect the same measured structures that are useful for predicting the label. In chipless RFID, reader geometry can alter the visibility and relative structure of resonant features that also carry TagID information. Alignment or invariance is therefore not automatically beneficial. A method that suppresses geometry-associated variation may improve robustness, but if the removed variation overlaps with class-discriminative structure it may also reduce TagID

separability. Whether useful identity information can be separated from acquisition-geometry information is consequently an empirical question rather than an assumption of this dissertation.

Model selection introduces a second difficulty. Domain generalisation is intended to evaluate performance on an unseen target, yet hyper-parameters and checkpoints still need to be selected. Gulrajani and Lopez-Paz showed in the DomainBed benchmark that conclusions about domain-generalisation methods can depend strongly on how this selection is performed, and that properly tuned empirical risk minimisation remains a strong reference under consistent protocols [6]. This motivates the source-bound selection strategy used here: target-position performance is not used to choose the primary model, and specialised objectives are compared with empirical-risk controls constructed within the corresponding experimental branch.

A related but scientifically different question arises once labelled target data become available. Few-shot support, prototype-based classification, trainable readouts and encoder fine-tuning no longer ask whether the model generalises without seeing the target; they ask how much target-specific calibration changes held-condition performance. Prototype methods are especially useful in this context because they construct class representatives from limited labelled support and classify queries by their relation to those representatives [29]. More generally, separating the learned representation from the downstream readout provides a useful diagnostic distinction: a transfer failure may arise because transferable TagID structure has been lost in the representation, because the final decision rule no longer fits that representation at the target geometry, or because both effects occur together. These possibilities motivate the controlled diagnostics developed later in Chapters 4 and 5.

2.1.4 Research gap and study boundary

The literature reviewed above provides mature frequency-coded and depolarizing tag designs, increasingly capable machine-learning decoders, and a broad toolbox for learning under distribution shift. The studies most directly relevant to this dissertation, however, do not jointly provide three pieces of evidence required to assess transfer to a held reader geometry.

First, familiar-condition decoding and held-geometry transfer need to be separated under an evaluation in which repeated measurements of the same physical condition cannot cross the training/test boundary. Without that separation, a high classification score does not establish that the decoder can operate at a reader geometry excluded from development.

Second, potential domain-generalisation remedies need to be evaluated under source-bound model selection and against appropriately matched empirical-risk controls. Otherwise an apparent improvement can reflect differences in training protocol or target-informed selection rather than the intended intervention itself.

Third, a failed transfer result needs to be localised rather than reported only as a lower accuracy value. In a physical sensing system the limitation may already be present in the measured signal, may emerge or worsen in the learned representation, or may lie partly in the final readout. Distinguishing these levels is necessary before deciding whether the appropriate remedy is a new learning objective, limited target calibration, a different representation, or a change to the sensing system itself.

This dissertation addresses these questions using the depolarizing-tag measurement campaign as its principal corpus. The primary governed evaluation excludes P4 from preprocessing, fitting and source-only model selection and treats it as the held geometry. This protocol boundary should not be confused with a claim that P4 was historically untouched across the wider research programme: earlier related analysis had examined measurements from P4. The study is therefore interpreted as a controlled source-only evaluation within the final experimental protocol, not as a fully prospective experiment on a target that had never previously been observed.

The earlier Tyndall sensor-tag campaign [3] is retained only for historical baseline context in Chapter 4. Separately, an **additional external four-class chipless-RFID corpus** [32] is used only for questions that its different tag codebook can legitimately answer, such as scoped pipeline portability and representation-learning checks; it is not treated as a direct class-for-class transfer benchmark for the seven-class principal task. Other supporting analyses are likewise interpreted according to the information available to them. This access-aware separation establishes the methodological boundary for Chapter 3 and for the experimental results that follow.

2.2 Chipless-RFID Sensing from First Principles

The preceding section positioned the work within the chipless-RFID and domain-generalisation literature. This section now follows the specific sensing chain used in the principal dataset: how tag identity is encoded in a passive structure, how that structure is interrogated through a cross-polar measurement, how acquisition geometry can alter the resulting spectrum, and how the physical problem becomes a machine-learning task.

2.2.1 Spectral encoding and the physical tag

The tags studied in this dissertation encode identity directly in passive resonant geometry. A frequency-coded chipless tag contains several resonant elements whose electromagnetic responses depend principally on their physical dimensions and surrounding dielectric environment. When illuminated over a frequency band, each resonant element can contribute a local feature to the scattered response. The combination of these features forms a spectral signature from which the tag identity can be inferred.

The depolarizing tag family used in the source measurement campaign provides a concrete example. The physical Tag 111 introduced in Figure 2 is a representative member of the seven-tag classification set and contains four slanted resonant elements. Elements L1–L3 provide the three identity-bearing components, while Lc provides a fixed reference resonance [4]. Different configurations of the identity-bearing elements therefore produce different spectral patterns and form the tag codebook.

Crucially, TagID is not attached to the measurement as an independent digital message; it is expressed through electromagnetic structure in the measured spectrum. Any physical factor that changes the visibility or relative form of those resonant features can therefore also change the input eventually presented to the classifier.

2.2.2 Cross-polar readout and the measured spectrum

Because the tag is a passive scatterer, its response is measured together with electromagnetic contributions from the surrounding scene and measurement path. In a co-polar configuration, reflections from the supporting structure and other objects can contribute strongly to the received signal. The tag family used here addresses part of this problem through depolarisation.

The resonant elements are oriented so that part of the incident energy is re-radiated into the orthogonal polarisation. A receiver oriented orthogonally to the transmitted polarisation can therefore suppress a substantial component of the co-polar background while retaining the tag's cross-polar response [12], [13]. This does not isolate the tag perfectly from its environment, but it improves the observability of the resonant structure on which identification depends.

Figure 3 illustrates the bistatic cross-polar measurement principle used in the source study. A vector network analyser drives the transmitting antenna over a swept frequency range, the tag scatters the incident field, and an orthogonally polarised receiving antenna measures the cross-polar response. The resulting measurement is subsequently processed into the spectral representation retained by the dataset.

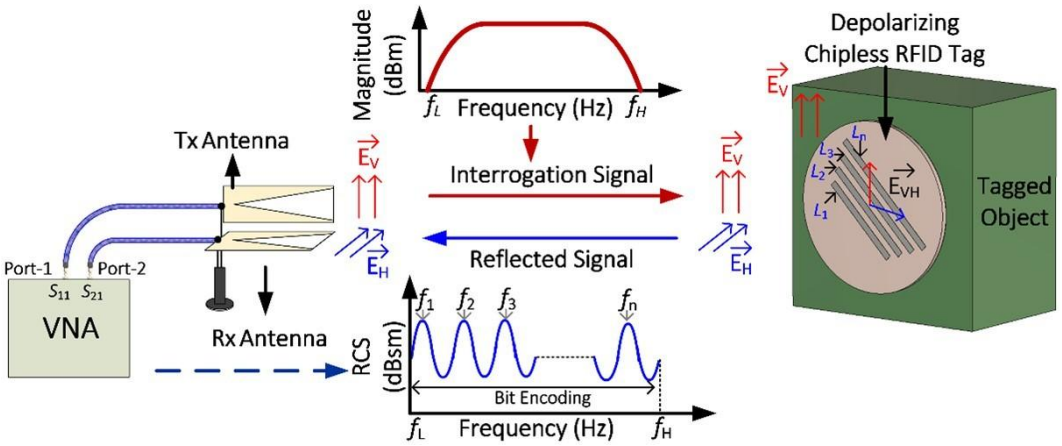

Figure 3. Depolarizing chipless-RFID cross-polar measurement principle. The transmitting antenna illuminates the passive tag and the orthogonally polarised receiver captures the frequency-dependent cross-polar response. Reproduced from the verified source-study configuration [4].

The principal dataset retains a cross-polar radar-cross-section magnitude spectrum derived from this measurement chain. Plotted against frequency, the spectrum contains the resonant structure through which the physical tag is identified. Figure 4 shows a representative simulated and measured response for Tag 111 from the source study [4].

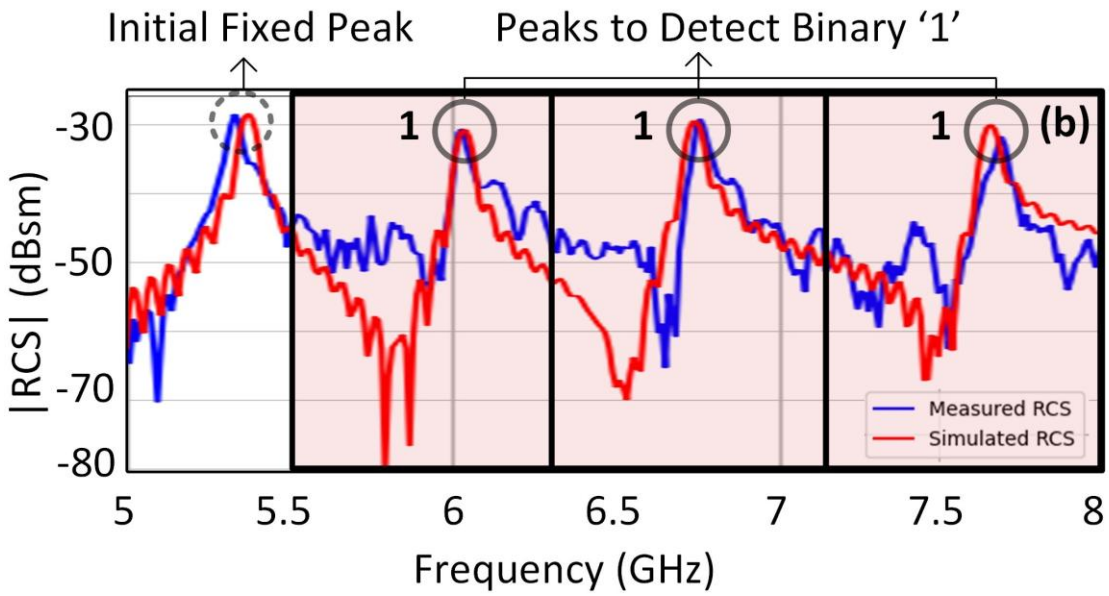

Figure 4. Representative simulated and measured cross-polar response of Tag 111. The spectrum shows the fixed reference resonance and the identity-bearing resonant features. Reproduced from [4].

Three properties of the retained measurement are particularly important.

First, it is **magnitude-only**. Complex phase information is not available to the classifier, so information carried only in phase cannot contribute to the learned decision.

Second, it is **one-dimensional**. The learning system receives a frequency-dependent spectral vector rather than a spatial image of the physical tag or its surrounding scene.

Third, it is **geometry-dependent**. Measuring the same physical tag from a different reader distance or acquisition angle can produce a different spectral vector. The classification problem therefore requires TagID to be recovered from a signal that depends not only on the tag itself but also on how that tag is observed.

The exact sampled signal, preprocessing operations and representation-specific transformations used by the individual experiments are defined in Chapter 3.

2.2.3 How acquisition geometry enters the signature

Reader position is represented in the principal dataset by two acquisition factors: reader-to-tag distance and acquisition angle. These variables need not alter the spectrum through one simple mechanism. Several physically plausible effects can contribute simultaneously, and separating them conceptually is useful before the empirical diagnostics are introduced.

Amplitude and signal visibility. Increasing reader distance is generally expected to reduce the received signal level under otherwise comparable conditions. Shallow resonant features may consequently become less visible relative to measurement variability and the effective noise floor. The magnitude of this effect depends on the complete antenna–tag–environment geometry and is therefore tested empirically rather than assumed from distance alone.

Illumination across the finite tag area. Changing reader distance also changes the spatial field distribution incident on the tag. Different ranges can alter how uniformly the resonant elements are illuminated and how strongly individual elements are excited relative to one another. A distance change can therefore affect relative spectral structure rather than acting only as a global amplitude scaling.

Polarisation efficiency. The depolarizing response depends on the relationship between the incident electric field, the orientation of the resonant elements and the orthogonal receive channel. Changing reader orientation changes that relationship and can therefore alter how much of the tag response appears in the measured cross-polar component. This provides a physically plausible route by which acquisition angle can influence resonance visibility.

Interaction with material loading. The tag substrate and the material beneath the tag contribute to the electromagnetic loading of the resonant structure and can alter the measured resonance pattern. Their effects may also interact with illumination geometry. Surface condition and reader position remain distinct experimental factors, but their effects on the measured response need not combine additively.

These mechanisms should be interpreted as plausible physical routes rather than as causal mechanisms established by the present dataset. The experimental campaign varies

acquisition conditions but does not independently intervene on each electromagnetic mechanism. Later factor-aware analyses therefore test associations between geometry and recognition behaviour without treating those associations as proof of a unique physical cause.

The important machine-learning consequence is nevertheless clear. Geometry-associated variation can affect spectral structure that is also useful for distinguishing TagIDs. The resulting question is not simply whether geometry information can be removed, but whether it can be reduced **without suppressing spectral information that also contributes to TagID discrimination**. This distinction motivates the representation diagnostics and source-only domain-generalisation experiments developed later in the dissertation.

2.2.4 Formalising the identification problem

It is useful to separate the physical measurement from the representation subsequently presented to the classifier.

Let the retained measured spectrum be (Eq. (1))

$$r = \mathcal{M}(y, e, s, d, \varepsilon) \tag{1}$$

where y denotes the physical TagID, e denotes the relative-permittivity condition, s denotes the mounting-surface condition, d denotes the reader position or domain, and ε represents repeat-level measurement variability and other unmodelled effects. The operator $\mathcal{M}$ is used only as a conceptual description of the physical measurement process; no explicit electromagnetic inverse model is assumed.

The measured spectrum is then transformed into the representation used by a particular learning pipeline (Eq. (2)):

$$x = T(r) \tag{2}$$

where T denotes the relevant preprocessing and representation transformation. Depending on the experiment, this may preserve the retained spectrum directly or transform it, for example through first differencing and source-derived standardisation. Consequently (Eq. (3)),

$$r \in \mathbb{R}^{N_r}, \quad x \in \mathbb{R}^{N_x} \tag{3}$$

and N_x need not equal N_r. The exact dimensions and transformations are defined in Chapter 3 rather than being assumed at this stage.

The classification problem is then to learn (Eq. (4))

$$f_\theta(x) \rightarrow y \tag{4}$$

so that TagID can be recovered from the representation derived from the measured spectrum.

For the primary Strict-DG experiment (Eq. (5)),

$$\mathcal{D}_S = \{\text{P1}, \text{P2}, \text{P3}\} \tag{5}$$

defines the source domains, while (Eq. (6))

$$\mathcal{D}_T = \{\text{P4}\} \tag{6}$$

defines the held target domain. The seven TagID classes remain unchanged across the four positions. What changes is the measurement process through which those identities are observed.

A successful source-only model must therefore learn TagID-discriminative structure from P1–P3 that remains useful when the measurement geometry changes to P4. It is not required to reconstruct or invert the underlying electromagnetic measurement operator. The engineering question is narrower and directly testable:

Does the learned decision remain predictive when the measurement process changes to a reader geometry that did not participate in model development?

This formulation also fixes the physical grouping used in the next chapter. A physical condition is defined by a fixed tuple

$$(\text{y}, \text{e}, \text{s}, \text{d})$$

while repeated measurements under that condition differ through repeat-level variability represented conceptually by ε_i. Chapter 3 formalises these measurements as condition blocks, specifies the corresponding grouping rules and defines which information each evaluation regime is permitted to access.

2.3 What the Source Study Established — and What It Left Open

The principal dataset used in this dissertation was originally acquired for a different but complementary purpose. The source study demonstrated high TagID identification performance on measurements drawn from a deliberately varied campaign that included changes in relative permittivity, mounting surface, reader distance and reader angle [4]. Its objective was to establish whether depolarizing chipless-RFID tags remained identifiable within that varied acquisition population.

That result established an important premise for the present work: the measured spectra contain learnable TagID information when model development and evaluation draw from the same mixed acquisition population. It did not, however, establish whether a model

developed without one reader geometry could transfer to that geometry when it was encountered only at evaluation.

This dissertation begins at that unresolved boundary. It changes the generalisation test from another measurement drawn from the pooled campaign to an entire held reader position. The primary evaluation therefore uses condition-block-grouped splits and separates source-only model development from experiments that are explicitly permitted to access target-domain information. Table 2 summarises the resulting change in scientific question.

Table 2. The source study and the present dissertation ask different questions of the same principal measurement campaign.

Dimension	Source study [4]	This dissertation
Primary question	Can a deep model identify depolarizing tags when trained and evaluated on measurements drawn from a deliberately varied acquisition campaign?	Can a model developed using only P1–P3 identify the same tags at the held P4 reader geometry?
Evaluation structure	Individual spectra are evaluated within the pooled acquisition population, with reader positions belonging to the same overall development/evaluation campaign.	Physical condition blocks are kept intact, and an entire reader position is reserved as the held target geometry.
Target access	No reader geometry is reserved as an unseen source-only target.	P4 does not enter the primary governed source-only development path.
Scientific interpretation	Evidence that TagID is learnable under familiar, mixed physical conditions.	A source-only estimate of held-geometry transfer, followed by signal–representation–readout diagnosis and controlled remedy tests.

Chapter 3 now translates this physical and scientific distinction into the concrete dataset grain, condition-block definition, source-only selection procedure, information-access rules and evaluation statistics used throughout the remainder of the dissertation.

3 Methodology and Evaluation Design

Chapter 2 separated the physical measurement process $r = \mathcal{M}(y, e, s, d, \varepsilon)$ from the representation $x = T(r)$ presented to the learning system. This chapter turns that abstraction into the concrete experimental design. Section 3.1 defines the principal measurement corpus, the spatial domains, the physical condition block and the signal representations. Section 3.2 defines what information each study is permitted to use and how the canonical source-only recipe is selected. Section 3.3 then specifies the metrics, inferential units, uncertainty procedures and learning-validity controls used to interpret the results.

3.1 Dataset and Measurement Design

3.1.1 Principal measurement campaign

Table 3 summarises the cross-polar depolarizing-tag measurement campaign reported in [4]. The dataset contains 12,600 spectral signatures from seven TagIDs. Each tag was measured at three relative-permittivity classes, under three repository-defined surface conditions (A1–A3), and at four reader positions. Every TagID–permittivity–surface–position combination was measured fifty times. A1–A3 are repository-defined surface-condition labels; their precise physical mapping is not used for the principal inference.

Table 3. Structure of the principal cross-polar depolarizing-tag dataset.

Property	Principal dataset
Measurement rows	12,600
TagIDs	7
Relative-permittivity classes	3
Surface conditions	3 (repository labels A1–A3)
Reader positions	4 (P1–P4)
Repeated sweeps per physical condition	50
Condition blocks per position	63 (7 × 3 × 3)
Total physical condition blocks	252
Supplied classification representation	281-point cross-polar RCS magnitude spectrum
Primary role	Principal corpus for Chapters 4–5

The measurements were collected using the automated robot-assisted acquisition system developed in the source study [4] and shown in Figure 5. The system combined controlled material and mounting conditions with repeatable presentation of the tags to a VNA-based cross-polar reader, allowing the complete factorial campaign to be acquired systematically.

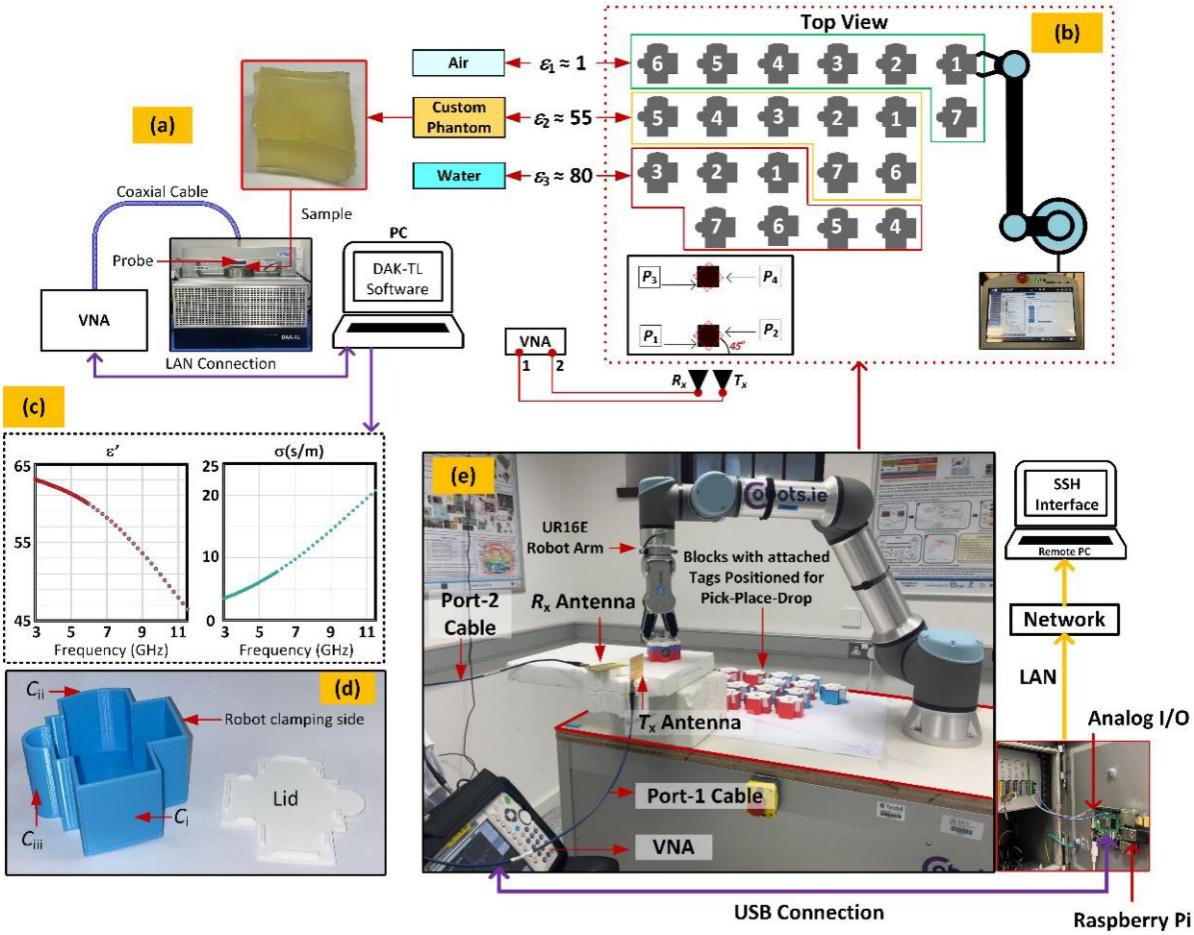

Figure 5. Automated acquisition system used for the principal depolarizing-tag campaign. Robot-assisted measurement system from the source study, showing the measurement structures, automated tag presentation and VNA-based acquisition arrangement. Reproduced from [4].

The resulting corpus is therefore a complete, deliberately structured measurement campaign rather than a collection of independent observations. At every reader position, all seven TagIDs occur under all nine permittivity–surface combinations, with fifty repeated sweeps of each physical condition. The dissertation analyses these measurements; it did not manufacture the tags or repeat the original acquisition campaign.

A secondary Tyndall sensor-tag campaign [3] is retained only for the historical baseline context of Chapter 4 and does not enter the principal Strict-DG development path. That dataset contains 9,600 monostatic, single-antenna, S11-derived RCS signatures from eight TagIDs, three capacitance states (0.1, 0.3 and 0.8 pF), four reader positions, five mounting/deformation cases and twenty repeated readings per physical condition. Its factorial structure therefore contains

$$8 \times 3 \times 4 \times 5 = 480$$

physical condition blocks, or 120 per position, with

$$480 \times 20 = 9600$$

measurement signatures in total. The source measurements contain 700 frequency samples over 3.1–10.6 GHz.

The P1–P4 labels in that earlier campaign are dataset-specific. They use 200- and 300-mm read ranges with 0° and 45° configurations, whereas the principal depolarizing-tag dataset uses 50 and 150 mm. Identically named position labels across the two campaigns therefore denote analogous experimental categories rather than identical physical reader locations, and no cross-dataset positional equivalence is assumed.

3.1.2 Spatial domain design: P1–P4

The domain shift studied in this dissertation is defined by the geometry of the reader relative to the tag. The four positions form a 2 × 2 design over reader distance and observation angle. Table 4 and Figure 6 summarise this geometry.

Table 4. Reader positions defining the principal domain-generalisation task.

Position	Distance	Angle	Role in this dissertation
P1	50 mm	0°	Source domain
P2	50 mm	45°	Source domain; shares the target angle
P3	150 mm	0°	Source domain; shares the target distance
P4	150 mm	45°	Held target; unobserved joint factor combination

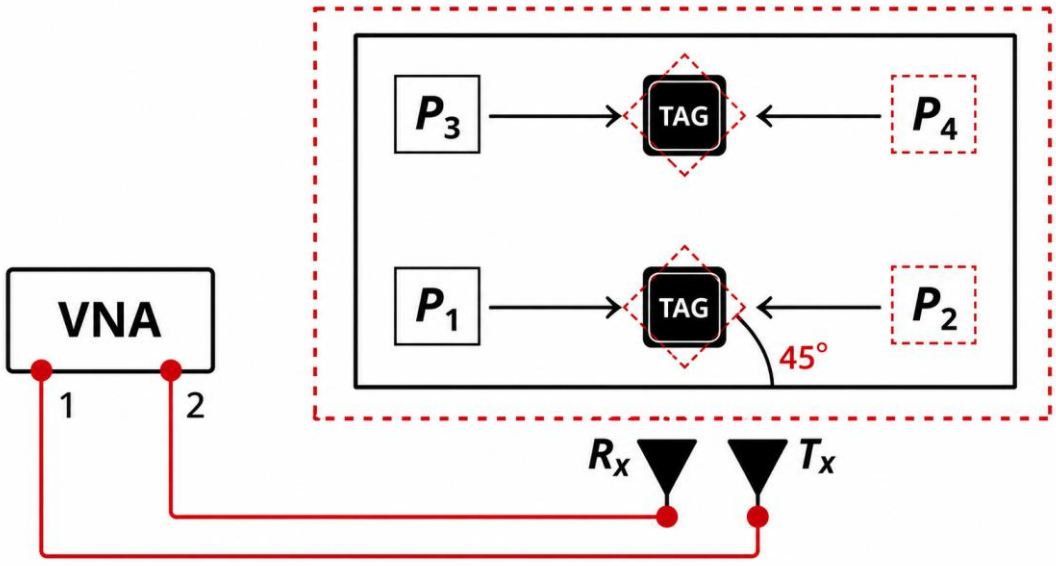

Figure 6. P1–P4 geometry defining the principal domain-generalisation problem. P1/P2 use the 50-mm distance, P3/P4 use 150 mm, P1/P3 use 0°, and P2/P4 use 45°. P1–P3 form the source domains and P4 is the held target. Adapted from [4]; schematic not to scale.

This structure defines the central transfer problem. P1, P2 and P3 are available during source-domain development, whereas P4 is reserved as the held target in the canonical Strict-DG evaluation. Each target factor level has therefore been observed individually in the source data: P2 supplies the 45° condition and P3 supplies the 150-mm condition. Their joint occurrence at 150 mm and 45°, however, has not been observed during source development.

Transfer to P4 consequently tests generalisation to an unobserved joint factor combination rather than interpolation among previously observed geometries. Its role follows from this information boundary; P4 is not assumed to be intrinsically the hardest position. Because the design contains only one such held joint corner, the resulting conclusions describe this specific geometry shift rather than position shift in general.

3.1.3 Condition blocks and the physical evaluation unit

Chapter 2 represented the retained measured spectrum as (Eq. (7))

$$r = \mathcal{M}(y, e, s, d, \varepsilon) \tag{7}$$

with the representation supplied to a learning pipeline subsequently written as (Eq. (8))

$$x = T(r) \tag{8}$$

For a fixed physical tuple (y, e, s, d), the principal campaign records fifty repeated measurements. A **condition block** is therefore defined as (Eq. (9))

$$B_{y,e,s,d} = \{\, r_{y,e,s,d,\varepsilon_i} : i = 1, \ldots, 50 \,\} \tag{9}$$

The factors y, e, s and d define the nominal physical condition, while ε_i represents repeat-to-repeat acquisition variability under that condition. The fifty rows within a block are therefore repeated measurements under one nominal physical state, not fifty independent environmental conditions or independent acquisition campaigns.

Every transformed measurement inherits the condition-block identity of its underlying measured spectrum. The grouping rule is defined before the choice of RAW, first-difference or learned representation and is not altered by subsequent preprocessing.

At each reader position,

$$7\ \text{TagIDs} \times 3\ \text{permittivity classes} \times 3\ \text{surface conditions} = 63$$

condition blocks are present. With fifty sweeps per block, each position contains

$$63 \times 50 = 3150$$

Across all four positions, the principal dataset therefore contains 252 physical condition blocks and 12,600 measurement rows.

A separate repository audit identified **4,299 unique hashes of the 281-point processed spectra** among those 12,600 rows. This quantity is distinct from the 252 physical condition blocks. Signal hashing identifies exact numerical duplication of the supplied spectra; it does not define statistical independence. Two spectra may have different hashes while still being highly related repeated measurements of the same physical condition. Conversely, repeated rows with exactly the same 281 values share the same hash.

For this reason, scientific grouping is defined by the physical metadata (y, e, s, d), not by signal hashes. In the principal held-condition and cross-position evaluations, all fifty repeats belonging to one condition block remain together whenever a train/test boundary is constructed. Experiments that intentionally permit shared physical conditions, such as the explicitly labelled within-condition interpolation diagnostic, are treated as a separate evaluation regime rather than as exceptions hidden inside the principal protocol.

The block structure also provides a useful physical reference scale for interpreting row-level Accuracy. Within a single position-level evaluation, changing all fifty predictions of one block from incorrect to correct changes row Accuracy by

$$50 / 3150 \approx 0.0159$$

This does not define a statistical significance threshold. It is instead a physical reference scale for judging whether a small numerical difference corresponds to coherent recovery of an entire physical condition or only to scattered row-level changes.

3.1.4 Signal representations and preprocessing

The classification experiments begin from the supplied 281-point cross-polar RCS magnitude spectrum. This representation is referred to as **RAW** in the experimental nomenclature, although it is a cropped measurement representation rather than the complete original instrument sweep.

Clarification from the original study team confirmed that these 281 samples correspond to the **5–8 GHz frequency range cropped from the original 700-point sweep** (personal communication, July 2026). The intended crop band is therefore established, but the authoritative sample-by-sample frequency coordinates and exact original crop indices are not available. The dissertation treats them as the ordered spectral samples of the supplied 5–8 GHz encoding-band representation. The principal experiments operate directly on these supplied samples rather than attempting to reconstruct the discarded portions of the original sweep.

Three representations are central to the main analysis.

The first is the supplied spectrum (Eq. (10)),

$$x_{\text{raw}} \in \mathbb{R}^{281} \tag{10}$$

It preserves the amplitude structure available in the deposited classification corpus and provides the least transformed reference in the representation diagnostics of Chapter 4. The supplied representation contains magnitude information; complex phase and additional polarimetric channels are not available.

The second is its discrete first difference (Eq. (11)),

$$x_{\text{diff,i}} = x_{\text{raw,i+1}} - x_{\text{raw,i}}, \quad i = 1, \ldots, 280 \tag{11}$$

which gives (Eq. (12))

$$x_{\text{diff}} \in \mathbb{R}^{280} \tag{12}$$

First differencing reduces sensitivity to slowly varying baseline structure and emphasises local spectral changes. It is one of the source-only candidate representations evaluated in Chapter 4 and is the representation selected by the source-side procedure for the canonical Strict-DG pipeline.

For the canonical source-only pipeline, feature-wise standardisation parameters are estimated only from the permitted source training data and are then applied unchanged to held data. P4 is never used to estimate preprocessing state.

The third representation is the learned embedding (Eq. (13)),

$$z \in \mathbb{R}^{256} \tag{13}$$

defined as the penultimate output of the trained convolutional encoder. It is used in the representation, readout and target-access diagnostics to determine whether TagID discrimination available in the measured spectrum remains accessible after learned encoding.

A separate historical 512-point interpolated representation is retained only where required by the earlier baseline and selected physical diagnostics in Chapter 4. It is not an additional measured modality and is not the canonical Strict-DG input.

Direct comparison of RAW, first difference and z later provides the principal representation-level diagnostic of whether discrimination loss is already present in the measured signal or emerges additionally during learned encoding.

3.2 Information Boundaries and Evaluation Regimes

The central methodological rule is simple: **a result can answer only a question permitted by the information used to produce it.**

This distinction is particularly important in domain generalisation. A score on a held target can be interpreted as source-only transfer only if the target has not influenced preprocessing, fitting, representation or hyper-parameter selection, checkpoint selection, or model promotion. Once target labels or outcomes influence one of those decisions, the experiment may remain scientifically useful, but it answers a different question.

As discussed in Section 2.1, model-selection protocol can materially change the interpretation of domain-generalisation results [6]. The studies in this dissertation are therefore organised into explicit information-access classes. Each class records what information is permitted during fitting and selection and thereby determines the claim that may legitimately be attached to the resulting number.

The same numerical metric can therefore support a different scientific claim under a different information-access regime.

Principal split structures

Table 5 summarises the principal split structures. The row counts describe the amount of data available to each partition; their scientific meaning is determined jointly by the grouping and information-access rules.

Table 5. Split structures used for the principal depolarizing-tag corpus.

Split structure	Train	Validation	Test / held evaluation
Condition-grouped familiar-condition split	7,350	2,450	2,800
Strict P4 outer split	7,350	2,100	3,150
Source-selector fold within P1–P3	4,200	2,100	3,150 held-source rows
Leave-surface-case diagnostic	6,650	1,750	4,200

The familiar-condition split keeps physical condition blocks intact while allowing blocks from the same overall acquisition population to appear across development and test partitions. The strict P4 outer split instead reserves all 63 P4 condition blocks, or 3,150 rows, for the final held-geometry evaluation.

Canonical source-only model selection

The canonical source-only selector is itself evaluated entirely within P1–P3. Four candidate recipes are compared: raw-spectrum empirical risk minimisation (C0), first-difference empirical risk minimisation (C1), a source-only nearest-class-mean readout (C2), and CORAL-regularised empirical risk minimisation (C3).

Each candidate is evaluated under three leave-one-source-position-out folds. In each fold, one of P1, P2 or P3 acts as the held source position, while the remaining two source

positions provide 4,200 inner-training and 2,100 inner-validation rows. Five training seeds (42–46) are used in every fold. The resulting selector therefore contains

$$4 \text{ candidates} \times 3 \text{ held-source folds} \times 5 \text{ seeds} = 60$$

source-side candidate–fold–seed evaluations.

Candidates are ranked using the source-only lexicographic selector, whose primary criterion is worst held-source Macro-F1 across P1–P3. C1 wins on this primary criterion, so later tie-breaks do not affect the canonical choice. This deliberately conservative criterion favours the recipe whose weakest source-held result is strongest rather than the recipe with the best average or the best P4 result. No P4 measurement, label, metric or checkpoint outcome participates in this promotion decision. After selection, the complete recipe is frozen before the final governed evaluation on P4. Principal protocol and optimisation settings are summarised in Appendix B; complete implementation specifications are retained in the frozen release records.

Information-access classes

Table 6 formalises the information-access taxonomy. The five broad bands correspond to the reader-oriented map introduced in Chapter 1; the more detailed rows retain the exact scientific meaning required by the individual studies.

Table 6. Information-access classes used throughout the dissertation. The broad bands correspond to Figure 1; each detailed class retains its own fitting, selection and interpretation rules.

Broad information band	Detailed access class	Information permitted in fitting	Information permitted in selection	Scientific question answered
Familiar / source-internal	Familiar / mixed-condition historical	Mixed campaign partitions; P4 may occur as an ordinary member of the pooled data	Historical validation within the mixed campaign	Is TagID learnable when training and test measurements come from a familiar acquisition population?
	Source-only diagnostic	P1–P3 only	P1–P3 only	Can source evidence justify a development decision or reveal transfer risk without using P4 in that analysis?
Strict source-only	Strict-DG	P1–P3 only; P4 excluded from preprocessing-state estimation and fitting	P1–P3 evidence only; no P4 representation, hyper-parameter, checkpoint or model selection	Does a rule developed from the source geometries transfer to held P4?
	External SSL / source-only DG	Principal depolarizing-tag source data plus permitted unlabelled external data	Source evidence only; no P4 labels or outcomes	Does external self-supervision alter P4 transfer while the target remains unavailable to development?

Broad information band	Detailed access class	Information permitted in fitting	Information permitted in selection	Scientific question answered
Post-freeze diagnostic	Target-domain diagnostic	P4 may train or evaluate a diagnostic only as explicitly declared	Only as declared by the diagnostic protocol	Where is the transfer failure localised after the canonical recipe has been fixed? Not a final DG score.
Target-assisted	Few-shot / target-assisted	Labelled P4 support is permitted under the declared protocol	Only as explicitly declared	What changes when controlled labelled target information is authorised?
	Within-condition interpolation	P4 partitions may share physical conditions as explicitly declared	Target partitions as declared	How well can repeated measurements of already represented physical conditions be interpolated?
Retrospective / supporting	Retrospective target-informed	Target features may be evaluated through the declared historical carrier	Full P4 outcomes may influence a retrospective decision	What can hindsight reveal about representation or readout sensitivity within that historical lineage?
	External portability	External corpus only	External corpus only	Does the analysis pipeline operate successfully within the independent external corpus?
	Physical analysis / simulation	Diagnostic measurements or simulation data as declared	Not used as a Strict-DG selector	What physical association or design hypothesis is supported?

The external-SSL branch remains target-free with respect to P4 but additionally uses explicitly permitted unlabelled data from an external corpus. It is therefore grouped visually with the strict target-free analyses while retaining its own detailed access definition.

The canonical Strict-DG computation is source-only at the execution level. P4 is excluded from preprocessing-state estimation, fitting, representation and hyper-parameter selection, checkpoint selection and model promotion, and is opened only after the complete source-selected recipe has been fixed for final evaluation.

One qualification is retained explicitly. The wider research programme had examined P4 in earlier lines of work. The canonical evaluation is therefore not presented as a historically untouched prospective-target experiment. Its narrower and verifiable claim is that the governed computation reported as Strict-DG excludes P4 from the decisions used to construct that result.

Because information access changes the scientific question, results from different access classes are not treated as entries in a common leaderboard. The following comparability rules govern their interpretation:

1. **The canonical Strict-DG result is the primary source-only transfer estimate.** Poor performance on P4 under this regime is a completed scientific result rather than a failed execution to be replaced by a more favourable target-informed score.

2. **Target-assisted results are not domain generalisation.** Few-shot support, larger calibration budgets, target-trained readouts and encoder fine-tuning deliberately authorise labelled P4 information. They quantify the effect of controlled target access and are interpreted only against protocol-matched anchors.

3. **Post-freeze diagnostics do not retroactively alter the strict result.** P4 may be used after the canonical source-only recipe has been frozen to diagnose measured-signal, representation or readout behaviour. Such diagnostics explain the fixed result; they do not participate in selecting it.

4. **Retrospective target-informed results are not prospective improvements.** Full P4 outcomes influence at least one retrospective decision, so these results cannot be promoted into the Strict-DG benchmark. In particular, the difference between the canonical strict result and the historical retrospective result cannot be attributed wholly to target access because representation lineage and readout family also differ.

5. **Within-condition interpolation is not held-condition transfer.** When target partitions share physical conditions, performance reflects reclassification of repeated measurements from already represented conditions rather than generalisation to unseen conditions.

6. **DG intervention branches require branch-matched controls.** IRM, DANN and GroupDRO use branch-specific training and aggregation procedures and are therefore interpreted through their matched effects relative to the corresponding ERM controls. Their absolute P4 scores do not form a valid cross-method leaderboard.

7. **External and physical studies retain their declared scope.** Within-corpus performance on an external dataset does not validate transfer of the seven-class principal depolarizing-tag model, and physical or simulation analyses do not become classification-improvement evidence unless such a link is separately validated.

These boundaries are enforced through the execution design rather than by narrative convention alone. Study-specific entry points, explicit data-access contracts, frozen prediction artefacts, controlled target-label access and independent metric checks were used to keep source-only, post-freeze diagnostic, target-assisted and retrospective branches distinct. Appendix D summarises the corresponding software controls and verification mechanisms; detailed artefact identities are retained in the frozen release manifest so that the main methodology can remain focused on the scientific design.

The following section defines the metrics, inferential units, uncertainty procedures and learning-validity controls used to interpret the resulting experiments.

3.3 Metrics, Uncertainty and Learning-Validity Controls

With the data structure and information boundaries established, this final methodological section defines how performance is measured, how uncertainty is quantified, and how basic learning functionality is distinguished from genuine generalisation failure.

3.3.1 Metrics and reference baselines

Classification performance is summarised primarily by **Accuracy** and **Macro-F1**. Accuracy is the fraction of classified items assigned the correct TagID and is retained because of its direct interpretation. The relevant evaluation item is stated by the individual study protocol. Row Accuracy, within-condition row Accuracy and Accuracy after aggregating a condition into one prediction are distinct quantities; condition-block resampling does not itself turn a row metric into a block-classification metric (Appendix C).

Accuracy alone, however, can conceal severe class imbalance in the predictions. The principal task contains seven balanced TagID classes. A classifier that predicts the same class for every sample therefore still obtains (Eq. (14))

$$\text{Accuracy} = \frac{1}{7} \approx 0.1429 \tag{14}$$

Its class-wise performance is much less favourable. For the single predicted class, precision is 1/7 and recall is 1, giving (Eq. (15))

$$F_1 = \frac{2(\frac{1}{7})(1)}{(\frac{1}{7})+1} = \frac{1}{4} \tag{15}$$

The remaining six classes have zero recall and therefore contribute an F1 score of zero. The resulting Macro-F1 is (Eq. (16))

$$\text{Macro-F1(single-class)} = (\tfrac{1}{7})(\tfrac{1}{4} + 0 + \cdots + 0) = \frac{1}{28} \approx 0.0357 \tag{16}$$

This illustrates why Accuracy and Macro-F1 answer complementary questions in this dataset. Balanced uniform random prediction has an expected Accuracy and Macro-F1 of approximately (Eq. (17))

$$\frac{1}{7} \approx 0.142857 \tag{17}$$

whereas single-class collapse retains the same Accuracy but produces a much lower Macro-F1. The dashed 1/7 reference used in later figures should therefore be interpreted as the

balanced uniform-random reference, not as a universal baseline for every possible predictor.

Macro-F1 is the unweighted mean of the class-wise F1 scores (Eq. (18)),

$$F_{1,\text{macro}} = \frac{1}{K}\sum_{k=1}^{K}\frac{2P_kR_k}{P_k+R_k}, \quad K = 7 \tag{18}$$

where P_k and R_k are the precision and recall of class k. When $P_k + R_k = 0$, the corresponding class F1 is defined as zero, consistent with the evaluation implementation.

Macro-F1 is used as the primary classification metric because zero-recall classes contribute zero to the class average and therefore expose failure modes that Accuracy can obscure. A Macro-F1 value below the uniform-random reference is not, by itself, proof of class collapse. It is interpreted together with **per-class recall** and the explicit **zero-recall count**. This combination makes it possible to distinguish diffuse classification error from the more severe case in which one or more TagIDs are effectively abandoned.

3.3.2 Units of inference and uncertainty

The repeated-measurement structure established in Section 3.1.3 determines the appropriate statistical unit. For analyses based on repeated measurements from the principal corpus, the condition block

$$B_{y,e,s,d}$$

is the default inferential unit rather than the individual measurement row. The fifty sweeps within a block are repeated observations of the same nominal physical condition and are therefore not treated as fifty independent experimental replicates.

This principle does not imply that every analysis in the dissertation has the same inferential unit. Different study families use the unit defined by their experimental design. Examples include condition blocks in the fixed-position, factorial and representation analyses; paired episodes in the historical few-shot study; and held-source events in the source-selection analysis. Where training-seed variability is examined, seeds provide an additional algorithmic source of variation rather than new physical acquisition campaigns. The relevant unit is stated explicitly for each result.

For the principal matched block-level analyses, uncertainty is estimated using a **10,000-replicate, TagID-stratified cluster bootstrap** over complete condition blocks. Where the comparison is paired, the pairing is preserved during resampling. This avoids artificially increasing the apparent sample size by resampling the fifty correlated rows within each physical condition independently.

The resulting 95% percentile intervals describe uncertainty over the observed condition-block population under the declared protocol. They should not be interpreted as confidence intervals over entirely new reader geometries. Other study families use their own protocol-defined resampling units; for example, the historical few-shot comparison uses a 10,000-resample paired-episode bootstrap.

Where variability across network initialisations is also scientifically relevant, a secondary sensitivity analysis additionally resamples the aligned training seeds. If the seeds-fixed and seed-inclusive intervals lead to different interpretations, both are reported and the more conservative interpretation governs the scientific claim.

Conventional parametric p-values are not used as the primary inferential device. The principal study contains only three source geometries, five training seeds and one held target geometry, and these quantities do not constitute a large set of independent physical replications. Effect estimates together with protocol-appropriate uncertainty intervals therefore provide a more direct description of the magnitude and stability of the evidence available at this experimental scale.

Dispersion across repeated training runs is reported as the **population standard deviation** (ddof = 0) unless stated otherwise. This quantity describes variation across the declared training seeds and should not be interpreted as a confidence interval over future physical deployments. Likewise, when a retrospective result is available from only one run, a reported standard deviation of zero is merely the arithmetic consequence of n = 1 and is never treated as evidence of algorithmic stability.

3.3.3 Learning-validity controls

A result close to the balanced uniform-random reference admits an important ambiguity. The held-condition task may genuinely be difficult, or the training pipeline may have failed to learn even a small labelled dataset. A separate **learning-validity control** is therefore used to distinguish basic learning functionality from held-condition generalisation.

As part of the fixed-position validation protocol, the architecture is asked at each reader position to fit a fixed **56-row true-label subset drawn from the corresponding training data**, without changing the primary hyper-parameters. The purpose of this test is not to demonstrate generalisation and not to require mathematically perfect memorisation. Instead, it checks whether the architecture, optimiser, loss and data path can drive training performance close to saturation on a deliberately small supervised problem.

The registered validity criterion requires both Accuracy and Macro-F1 to exceed 0.95. All four positions passed this gate. The recorded Accuracy values ranged from 0.9643 to 0.9821, with the criterion reached within 33–50 optimisation steps. The detailed records are:

1. P1: Accuracy 0.9643 in 37 steps;
2. P2: Accuracy 0.9821 in 38 steps;
3. P3: Accuracy 0.9821 in 33 steps;
4. P4: Accuracy 0.9643 in 50 steps.

The term *optimisation step* is used here in the sense recorded by the diagnostic implementation; detailed execution settings are retained in Appendix B.

The P4 learning-validity arm necessarily uses P4 labels, but it was executed only after the canonical Strict-DG recipe and source-only checkpoints had been frozen. It therefore does not participate in representation choice, hyper-parameter selection, checkpoint promotion or any other Strict-DG development decision. It belongs to the post-freeze diagnostic access band defined in Section 3.2.

Passing the learning-validity control establishes only that the basic supervised learning path is functional and that the architecture possesses sufficient capacity to fit this small true-label subset. It does **not** establish that the model should generalise across unseen condition blocks or across reader geometries. Its role is narrower: a weak P4 result can then be interpreted as a held-geometry generalisation failure under the declared protocol rather than being dismissed simply because the model never demonstrated an ability to learn its training labels.

With the physical data grain, information boundaries, source-only selection rule, metrics and uncertainty procedures now fixed, the dissertation can move from protocol to evidence. Chapter 4 first uses the historical baselines to motivate the protocol gap, then establishes the governed P1–P3 → P4 Strict-DG failure, and finally diagnoses that failure through within-position, acquisition-factor and representation analyses.

4 Establishing and Diagnosing the Transfer Failure

This chapter establishes the transfer failure under a governed source-only protocol and then localises where that failure arises. The argument follows a diagnostic funnel. Section 4.1 uses the historical results only to establish the motivating protocol gap. Section 4.2 then reports the canonical P1–P3 → P4 Strict-DG result under the information boundary defined in Chapter 3 and examines descriptive source-score support for the model choice. Section 4.3 tests whether P4 is intrinsically exceptional or instead belongs to a broader acquisition-dependent difficulty pattern. Section 4.4 finally localises the failure across the measured signal and learned representation. Chapter 5 then changes from diagnosis to intervention by testing source-only remedies and controlled target access.

4.1 Historical Baselines and Motivation for Strict Evaluation

The historical experiments provide the starting observation rather than the final transfer estimate. Their main value is to show that the apparent difficulty of the task changes sharply with the evaluation regime. Under familiar or mixed measurement conditions the principal depolarizing-tag dataset is readily learnable, whereas withholding an entire reader position produces a much lower score. These results predate the final Strict-DG governance machinery and are therefore retained as context only.

4.1.1 Historical grouped and leave-position behaviour

The historical deep-convolutional pipeline followed the source-study lineage [3], [4]. Table 7 retains three reference values: the familiar-condition result on the secondary co-polar sensor-tag dataset, the corresponding grouped result on the principal cross-polar depolarizing-tag dataset, and the historical leave-P4-out result on that principal dataset.

Table 7. Historical baseline results. These values motivate the transfer question but were not produced under the governed Strict-DG protocol of Section 4.2.

Protocol	Dataset	Accuracy
Deep CNN, condition-grouped random split	Co-polar sensor-tag dataset	0.9670
Deep CNN, condition-grouped random split	Cross-polar depolarizing-tag dataset	0.7344
Deep CNN, leave-P4-out	Cross-polar depolarizing-tag dataset	0.3185

For the principal dataset, grouped-random Accuracy falls from 0.7344 to 0.3185 when P4 is withheld. The training-set size is 7,350 rows in both cases, so this difference cannot be explained by a simple reduction in training data. It instead motivates the distinction between interpolation within a familiar measurement population and transfer to a reader geometry that is absent from development.

Condition grouping prevents repeated sweeps of the same physical condition from crossing the split, but it does not make the test distribution geometrically novel. Training and test blocks can still come from the same reader positions and from neighbouring material or surface conditions. The grouped-random result therefore describes interpolation within a familiar acquisition population rather than transfer to an unseen geometry.

The historical gap is consequently a motivation rather than a governed deployment estimate. Section 4.2 asks the stricter question in which P4 is excluded from fitting, preprocessing-state estimation and model selection.

4.1.2 Status of the historical evidence

The historical baselines were substantially reproduced in the current environment, including the headline grouped-random and leave-position means within the pre-specified tolerance. Minor run-level discrepancies are documented in Appendix D. Because these historical figures are used only as context, no principal conclusion of this dissertation depends on exact reproduction of every historical run.

4.2 Establishing the Transfer Failure: Strict Source-Only Domain Generalisation

Section 4.1 establishes a historical protocol gap. This section reports the primary governed result: whether a recipe developed only from P1–P3 can identify the same seven TagIDs at the unseen P4 geometry. It first defines the permitted information and candidate set, then examines descriptive source-score support for the model choice, and finally reports the held-P4 failure and its class-wise form.

4.2.1 Information boundary and pre-specified candidates

Features and labels from P1, P2 and P3 were available for model development. P4 information was excluded from fitting, preprocessing-state estimation, representation selection, architecture selection, hyper-parameter selection, early stopping, prototype construction, calibration choices and model promotion. P4 was opened only after the complete recipe and final source-trained models had been frozen.

One qualification remains important. The governed computation is source-only, but the wider research programme had previously examined P4 in other experimental lineages. The present experiment is therefore source-only at the computation and selection level, but it is not claimed to be a historically untouched-target study.

Table 8 summarises the four source-only candidates evaluated. The set is deliberately limited rather than exhaustive: it samples representative source-justified choices spanning a

neutral ERM baseline, a physically motivated representation change, an alternative readout and a classical alignment objective. More elaborate DG objectives are tested separately in Chapter 5 under their own branch-matched controls.

Table 8. The four pre-specified strict source-only candidates and how each was produced.

Candidate	Method	How it was produced
C0 neutral ERM	Empirical risk minimisation on the raw ordered spectrum; control	Reused the recorded neutral-baseline checkpoints and predictions
C1 first-difference ERM	Empirical risk minimisation on the 280-point first difference	Newly trained; selected by the source-only ranking and used for the final models
C2 source NCM readout	Source Euclidean nearest-class-mean readout	Reused the recorded encoder with a new source-only readout
C3 source CORAL	CORAL-regularised ERM under the recorded source-only protocol [22]	Newly trained

4.2.2 Source-only selection and selection regret

Selection used leave-one-position-out validation over the three source geometries. For each held source position, models were fitted on the remaining two and evaluated on the held one. Four candidates × three held-source folds × five seeds produced 60 source-selection units.

The source-only selector was lexicographic, with **worst held-source Macro-F1** as its primary criterion. This deliberately pessimistic first criterion reflects the deployment concern: a candidate that performs acceptably on average but collapses at one source geometry is a poor choice for transfer to an unseen one. C1 won on this primary criterion with a wide margin (Figure 7), so later tie-breaks did not affect the canonical choice. The final recipe uses first differences, source-only feature-wise standardisation, a GroupNorm 1-D convolutional network [28], AdamW [30], and five equally reported seeds. Final epoch counts for seeds 42–46 were 13, 13, 9, 10 and 9.

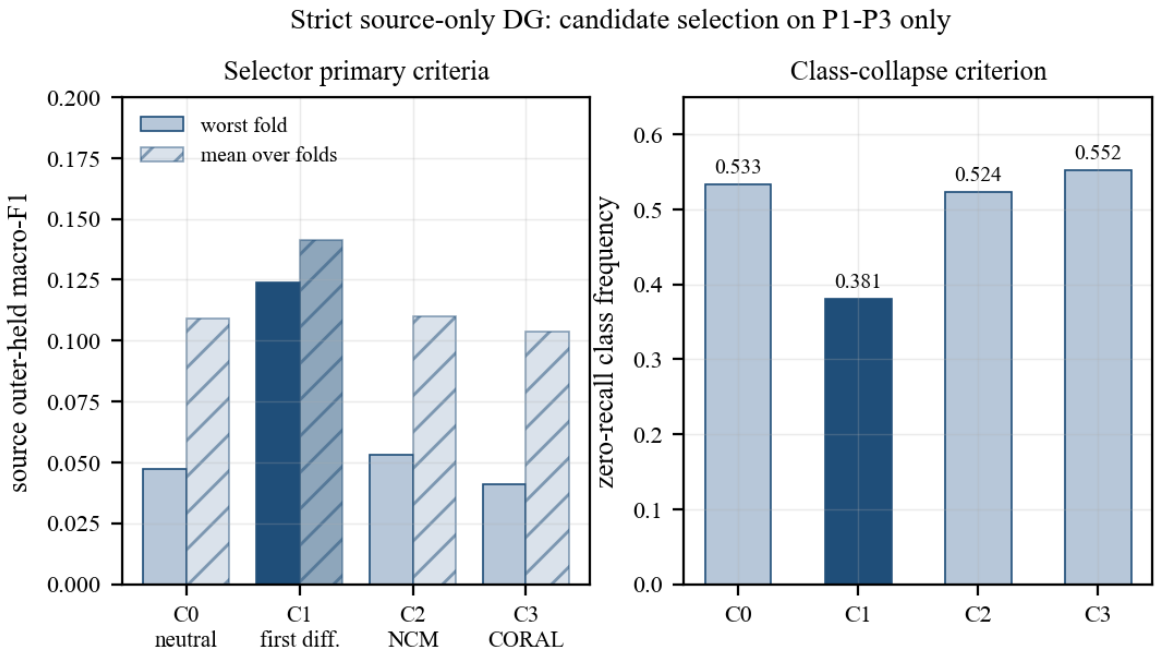

Figure 7. Source-side selection metrics for the four pre-specified candidates. The primary selection criterion is worst held-source Macro-F1; selection uses source-held evidence only and never consults P4.

For each held-source event, the selector was reapplied to stored scores for the other two source positions. This source-score reuse diagnostic is not a fresh nested development experiment in which the held position is excluded from all underlying model fitting. Selection regret is the held-position Macro-F1 of the oracle candidate minus that of the selected candidate, within the recorded candidate set (Table 9). P4 was not used in this diagnostic.

Table 9. Source-internal selection regret. The lexicographic selector is reapplied to stored score rows for the other two source positions; this is a source-score reuse diagnostic, not nested model development. Regret is oracle minus selected-candidate held-position Macro-F1.

Held position	Selector choice	Held-position oracle	Regret
P1	C1 first-difference	C1 first-difference	0.000000
P2	C1 first-difference	C3 source CORAL	0.040318
P3	C1 first-difference	C1 first-difference	0.000000

Mean regret was 0.013439, median regret was 0, and maximum regret was 0.040318; C1 was the held-source oracle in two of three events. These results provide descriptive support for C1 within the recorded four-candidate score set; they do not independently validate a nested held-source selection procedure. They do not establish that the selection rule is optimal for P4 or determine how much of the target failure could be avoided by a different source-only selection strategy. C3 CORAL was the held-source oracle at P2.

4.2.3 Held-P4 result and reconciliation with the historical baseline

After the source-only recipe was frozen, the five final models were evaluated once on P4, with the resulting held-target performance reported in Table 10.

Table 10. Strict source-only evaluation on the held P4 geometry. Selected recipe: first-difference GroupNorm 1-D CNN. The balanced seven-class reference is 1/7 = 0.1429.

Seed	Final epoch	Accuracy	Macro-F1
42	13	0.172698	0.123054
43	13	0.119365	0.096557
44	9	0.129206	0.100795
45	10	0.204127	0.144565
46	9	0.157778	0.096184
Mean ± population SD		0.156635 ± 0.030516	0.112231 ± 0.018956

Mean Accuracy was 0.1566 and mean Macro-F1 was 0.1122. The corresponding population standard deviations, 0.0305 and 0.0190, describe dispersion over the five training seeds rather than confidence intervals. The practical severity of the result is therefore interpreted primarily through its class-wise behaviour in Section 4.2.4 rather than through a formal test against the 1/7 reference.

The historical leave-P4-out Accuracy of 0.3185 in Section 4.1 and the canonical Strict-DG Accuracy of 0.1566 are **not** a matched before-and-after comparison. The historical branch used a different model and representation lineage, including the historical 512-point interpolated input, whereas the canonical branch uses the 280-point first-difference representation, source-only preprocessing state and the source-only selection rule defined above. Representation, architecture and validation lineage therefore change simultaneously. The numerical gap cannot be attributed specifically to stricter governance, source-side model selection or target leakage. The value 0.3185 is retained as historical motivation; 0.1566 is the governed transfer estimate on which the primary Strict-DG claim rests.

4.2.4 Class-collapse diagnosis

The aggregate score does not capture the most important form of the failure. The model does not weaken uniformly across all seven TagIDs. Instead, predictions concentrate on a small subset of classes, and every seed has at least one TagID with zero recall (Figure 8).

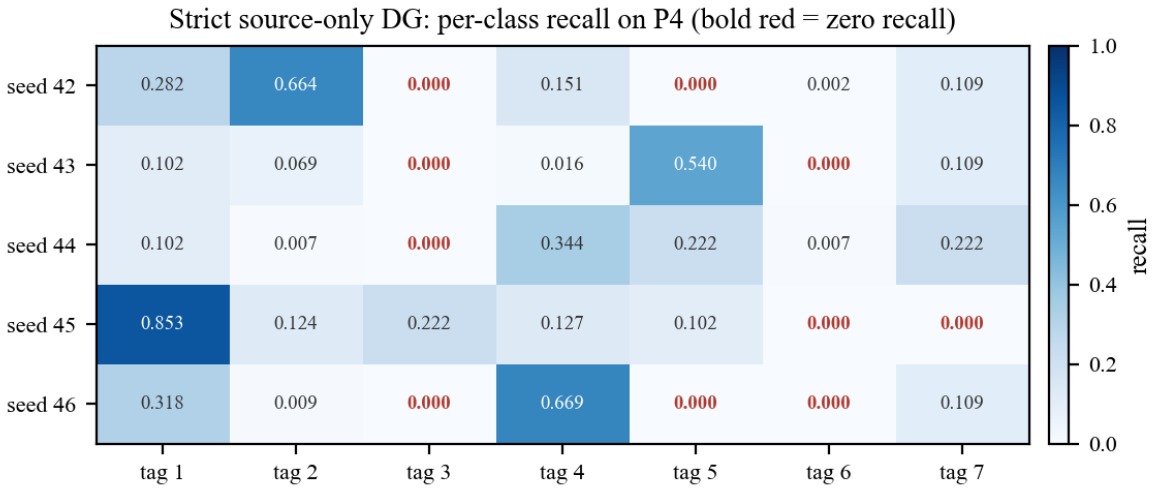

Figure 8. Per-class recall on the held P4 geometry by seed. Every seed has at least one zero-recall TagID.

This pattern illustrates why Macro-F1 was designated the primary metric in Section 3.3.1. Mean Accuracy remains close to the balanced seven-class reference, whereas Macro-F1 falls to 0.1122 because the errors are strongly class-asymmetric. Tags 3, 5, 6 and 7 each have at least one zero-recall seed. Tag 6 is the most severe pooled case, with only four correct classifications among 2,250 true Tag-6 test instances pooled across the five seeds.

The supported conclusion is therefore stronger than merely stating that the score is low: under the governed source-only protocol, C1 fails to achieve useful TagID transfer to P4 and its predictions collapse onto a restricted subset of classes. This does not establish that domain generalisation is impossible on chipless-RFID data, nor that an untested source-only representation could not perform better.

4.2.5 Evidence status and scope

The 60 source-selection units were replayed, independently re-scored and re-ranked from retained predictions rather than fully retrained. The five final Strict-DG models were retrained in the locked environment, and their stored P4 predictions and headline metrics reproduced exactly. Appendix D summarises verification status; detailed evidence identities are retained in the frozen release manifest.

The claim is additionally bounded by programme-level prior awareness of P4 outside the governed computation, one held target geometry and a four-candidate source-selection set. These limitations constrain the scope but do not alter the primary result: the declared source-only C1 pipeline does not transfer usefully to P4.

4.3 Diagnosing Acquisition Difficulty: Within-Position and Factorial Analysis

The Strict-DG experiment establishes that transfer to P4 fails. A remaining ambiguity is whether P4 is intrinsically exceptional or whether the failure reflects a broader acquisition-dependent reduction in TagID discriminability. The following diagnostics were executed after the canonical Strict-DG recipe had been frozen and therefore did not influence its development.

4.3.1 Within-position generalisation

The question is deliberately narrower than Strict-DG: given measurements of a tag under some physical conditions at one fixed reader geometry, can the same tag be recognised under unseen conditions at that same geometry?

Each position is treated as a separate dataset. Within one position, the 63 condition blocks are partitioned into three outer folds by a Latin-square assignment over the permittivity × surface grid. Training and held blocks therefore never share the same permittivity–surface cell for the same tag. Five seeds are run for each of the three folds at each of the four positions, giving 60 primary runs.

All four positions passed the separate learning-validity control described in Section 3.3.3; that control establishes a functioning supervised-learning path and sufficient small-set fitting capacity, not held-condition generalisation.

Table 11. Within-position generalisation to unseen condition blocks. Sixty runs were used (4 positions × 3 outer folds × 5 seeds). Accuracy and Macro-F1 score measurement rows; pooled-row Macro-F1 pools the held rows from all 63 blocks within each seed before averaging seeds. Intervals resample condition blocks. The balanced seven-class reference is 1/7 = 0.1429. Accuracy SD is the population SD over the 15 fold–seed runs at each position. The train-to-held gap is training row Accuracy minus held row Accuracy, averaged over those same runs.

Position	Held row Accuracy mean ± SD	95% interval (row Accuracy)	Pooled-row Macro-F1 [95% CI]	Train-to-held row Accuracy gap
P1	0.2688 ± 0.0775	[0.2147, 0.3257]	0.2348 [0.1788, 0.2846]	0.4247
P2	0.1499 ± 0.0564	[0.1081, 0.1948]	0.1194 [0.0771, 0.1581]	0.4737
P3	0.2389 ± 0.0799	[0.1902, 0.2927]	0.2164 [0.1631, 0.2649]	0.5344
P4	0.1711 ± 0.0573	[0.1300, 0.2154]	0.1322 [0.0929, 0.1699]	0.3037

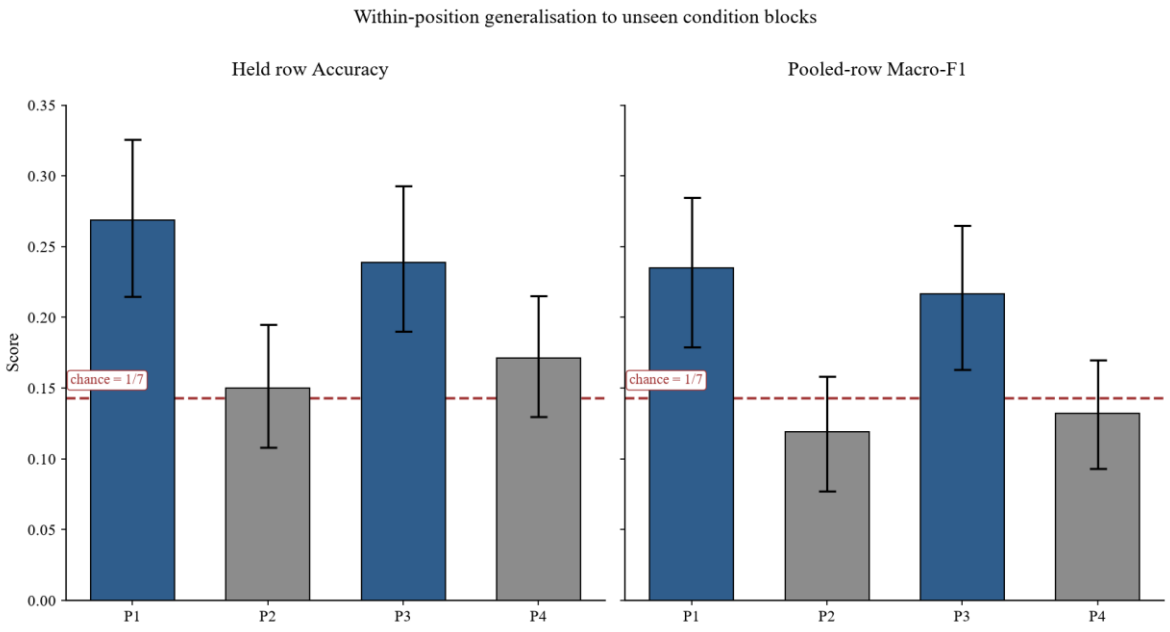

Figure 9. Within-position generalisation to unseen condition blocks. Both panels report row metrics; the Macro-F1 panel pools held rows across folds within each seed. Error bars show 95% condition-block cluster bootstrap intervals. P1/P3 are at 0° and P2/P4 at 45°; the dashed line denotes the balanced seven-class uniform-random reference (1/7 = 0.1429).

Table 11 and Figure 9 show that condition-general TagID identification is modest at every sampled position. Even at the strongest position, pooled-row Macro-F1 is 0.2348. The important comparison, however, is between P2 and P4. P4 is not uniquely weak under this matched protocol: P2 is numerically lower on the pooled-row Macro-F1, 0.1194 versus 0.1322, and on held Accuracy, 0.1499 versus 0.1711.

The more informative pattern is therefore not that P4 is the hardest position. P1 and P3 are the two 0° positions and form the stronger pair, whereas P2 and P4 are the two 45° positions and form the weaker pair. This motivates a synchronised factorial decomposition of angle and distance.

4.3.2 Synchronised 2 × 2 acquisition-factor analysis

The four positions are the corners of the two-factor design defined in Section 3.1.2:

- P1 = 50 mm, 0°;
- P2 = 50 mm, 45°;
- P3 = 150 mm, 0°;
- P4 = 150 mm, 45°.

To avoid confusion with the measurement operator $\mathcal{M}$ used in Chapters 2–3, let $m(P_i)$ denote the chosen performance metric at position P_i. The angle main effect averages the 0°→45° change at the two distances (Eq. (19)),

$$\Delta_{\text{angle}} = \frac{1}{2}[m(\text{P2}) - m(\text{P1}) + m(\text{P4}) - m(\text{P3})] \tag{19}$$

the distance main effect averages the 50→150 mm change at the two angles (Eq. (20)),

$$\Delta_{\text{distance}} = \frac{1}{2}[m(\text{P3}) - m(\text{P1}) + m(\text{P4}) - m(\text{P2})] \tag{20}$$

and the interaction compares the two angle effects (Eq. (21)),

$$\Delta_{\text{interaction}} = [m(\text{P4}) - m(\text{P3})] - [m(\text{P2}) - m(\text{P1})] \tag{21}$$

The primary endpoint is condition-block Accuracy after taking the highest mean logit within each of the 63 physical blocks, averaged across five seeds. Contrasts are paired within the same TagID–permittivity–surface condition. Primary uncertainty is a 10,000-replicate paired, tag-stratified block bootstrap with seeds fixed; seed-inclusive resampling is used as a sensitivity analysis where required.

An audit of the inherited runs found a validation-assignment coupling that prevented exact pairing across positions. The complete 60-run factorial experiment was therefore rerun with synchronised outer folds, seeds and position-independent validation assignment before inference. The amendment was recorded before synchronised training began; full provenance is given in Appendices B and D.

4.3.3 Acquisition-factor result

Table 12. Synchronised paired 2 × 2 acquisition-factor contrasts. Angle has a reliable adverse association; the overall distance main effect and the interaction are not confirmed.

Contrast	Accuracy estimate	95% interval	Supported?
Angle (45° − 0°)	−0.111111	[−0.160317, −0.063492]	Yes
Distance (150 − 50 mm)	−0.050794	[−0.106349, +0.007937]	No
Angle × distance interaction	+0.114286	[0.000000, +0.228571] seeds-fixed; [−0.028571, +0.260317] seed-inclusive	No

Corresponding pooled condition-block Macro-F1 contrasts are angle −0.109330 [−0.154894, −0.061699], distance −0.052759 [−0.105286, +0.001141], and interaction +0.116263 [+0.007935, +0.223512]. The interaction is nevertheless not confirmed because the governing seed-inclusive sensitivity interval crosses zero on the primary Accuracy endpoint.

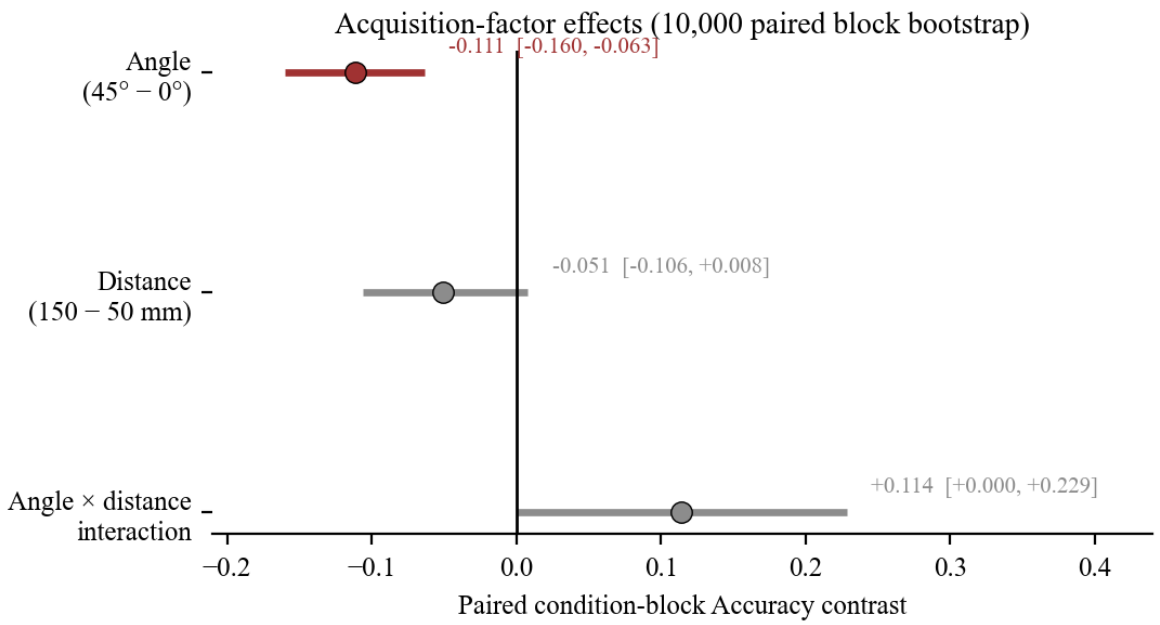

Figure 10. Synchronised acquisition-factor contrasts. Angle-associated degradation is supported; distance and the interaction are not confirmed under the governing uncertainty analysis. Error bars show 95% paired block-bootstrap intervals with seeds fixed; Table 12 also reports the governing seed-inclusive interaction interval.

Table 12 and Figure 10 show the clearest factor-level result: angle-associated degradation. Moving from 0° to 45° is associated with an approximately eleven-percentage-point reduction in condition-block Accuracy, with the interval remaining below zero. The corresponding Macro-F1 contrast is also negative.

The overall distance main effect is not confirmed. As a secondary contrast, the recorded distance effect (150 − 50 mm) at 0° is −0.107937 in Accuracy, with governing seed-inclusive 95% CI [−0.219048, +0.003175]; its seeds-fixed interval [−0.203175, −0.009524] is negative, but Accuracy support does not survive seed resampling. The corresponding pooled condition-block Macro-F1 effect is −0.110890, with seed-inclusive 95% CI [−0.213060, −0.008505], which remains negative. Thus support for this simple effect is endpoint-dependent. Distance cannot be dismissed as irrelevant, but the available data do not establish a reliable overall distance main effect.

The angle × distance interaction is also not confirmed. Its point estimate is positive, which would correspond to a smaller angle penalty at 150 mm than at 50 mm, but the governing seed-inclusive interval crosses zero. P2 and P4 also lie close to the classification floor. These floor-limited cells constrain the interpretability of the interaction, so the result should not be promoted into a physical angle-by-distance mechanism.

The reliable angle association is consistent with the orientation-dependent sensing mechanisms discussed in Section 2.2.3, including reduced cross-polar conversion efficiency

as one physically plausible explanation. The factorial result remains associational rather than proof of one isolated electromagnetic mechanism.

4.3.4 Supported interpretation

Taken together, the acquisition analysis shows that held-condition TagID identification is difficult across the sampled geometries, that 45° acquisition has a reliable adverse association with performance, and that neither a reliable overall distance main effect nor a physically interpretable angle-by-distance interaction is established.

P4 is therefore best understood as the unobserved joint 150-mm/45° geometry encountered within this broader acquisition-dependent difficulty pattern.

This localises part of the Strict-DG failure to the acquisition side of the problem. It does not yet establish what the source-trained encoder does to the discrimination that remains in the measurement itself. Section 4.4 addresses that question.

4.4 Signal and Representation Diagnostics

The transfer failure can in principle arise at three levels. First, the measured spectrum may already lose or distort TagID-discriminative structure under the changed acquisition geometry. Second, the learned encoder may fail to preserve the discrimination that remains. Third, the transferred readout may be mismatched to the target representation. This section addresses the first two levels; Chapter 5 later tests the readout and adaptation side under controlled target access.

4.4.1 Representation-comparison protocol

Three main representations of the same measurements are compared under an identical condition-block-disjoint linear-probe protocol at each position: Table 13 and Figure 11 present the comparison. In each position group in Figure 11, bars run from left to right as RAW, first difference and embedding, allowing series identification in monochrome.

- the raw 281-point spectrum;
- the 280-point first difference;
- the 256-dimensional penultimate embedding of the frozen source-only C1 encoder trained on P1–P3.

A fourth, pre-specified peak descriptor was also examined for physical interpretation, but it did not pass its extraction-validity gate and is therefore not used for quantitative representation comparison.

The same Latin-square fold assignment is used for every representation. The probe is a single linear softmax layer trained on the training blocks and evaluated on held blocks. The endpoint is Macro-F1 over the 63 out-of-fold condition-block majority predictions per position and seed.

The probe is deliberately linear. It asks how much TagID discrimination is linearly accessible from each representation; failure of the probe does not establish that all class information is absent.

4.4.2 RAW versus the learned representation at P4

Table 13. Condition-general TagID information by representation and position. Values are five-seed pooled 63-block Macro-F1. The final C1 encoder was trained on P1–P3; consequently P1–P3 are encoder-familiar diagnostics and only P4 is encoder-clean.

Position	RAW spectrum	First difference	Learned embedding
P1 (50 mm, 0°)	0.5192	0.3814	0.7632
P2 (50 mm, 45°)	0.2737	0.2481	0.7623
P3 (150 mm, 0°)	0.5363	0.3561	0.9095
P4 (150 mm, 45°)	0.2427	0.2359	0.1472

Table 14. Paired representation contrasts at the held P4 geometry, with 95% tag-stratified block-bootstrap intervals.

Contrast at P4	Estimate	95% interval
RAW spectrum − learned embedding	+0.0951	[+0.0173, +0.1747]
Learned embedding − first difference	−0.0876	[−0.1543, −0.0236]
RAW spectrum − first difference	+0.0075	[−0.0744, +0.0888]

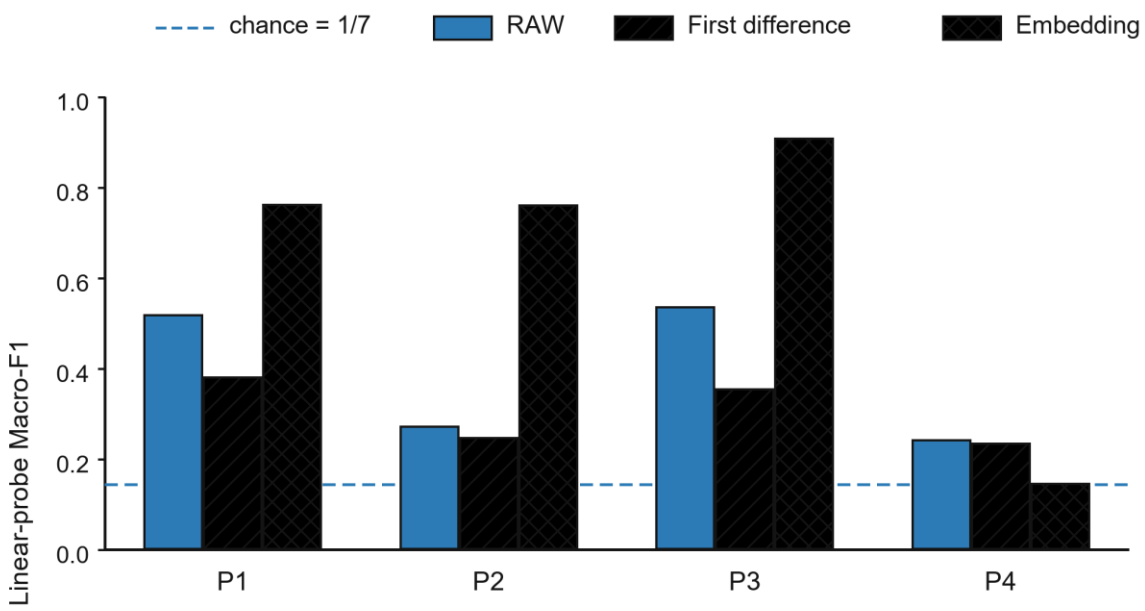

Figure 11. Linear-probe Macro-F1 for RAW, first difference and the learned embedding at P1–P4. P1–P3 are encoder-familiar/source-domain diagnostics; only P4 is encoder-clean. The dashed line denotes the balanced seven-class uniform-random reference (1/7 = 0.1429).

The paired contrast point estimates in Table 14 are means of the primary TagID-stratified condition-block bootstrap contrast distributions. They are therefore distinct estimands from, and need not equal the arithmetic differences between, the separately aggregated five-seed representation means in Table 13.

The P1–P3 embedding values are familiar/source-domain diagnostics because the final C1 encoder was trained on those positions. They are not unseen-position representation-generalisation estimates.

P4 is the encoder-clean comparison. There, RAW reaches 0.2427 Macro-F1, first difference reaches 0.2359, and the learned C1 embedding falls to 0.1472. RAW exceeds the learned embedding by 0.0951, with a 95% interval of [0.0173, 0.1747].

This result answers the representation-transfer question within the limits of the linear-probe diagnostic. At P4, the source-trained encoder does not preserve all of the TagID discrimination that remains linearly accessible in the raw spectrum. The supported claim is therefore a representation-dependent transfer limitation, not proof that TagID information has been destroyed, is absent from the embedding, or could not be recovered by another decoder.

The RAW probe value of 0.2427 is also numerically higher than the 0.1322 pooled Macro-F1 obtained by the separately trained within-position CNN in Section 4.3. These values arise from related but distinct diagnostic protocols and were not designed as a matched

architecture comparison. The difference should therefore not be interpreted as evidence that a linear classifier is intrinsically superior to a CNN at P4. Its narrower implication is that, under the representation-probe protocol, appreciable TagID discrimination remains linearly accessible in RAW. Specifically, the RAW probe scores one majority prediction per condition block, whereas the within-position CNN value scores pooled measurement rows.

4.4.3 Signal-level degradation precedes learned encoding

Table 15. Aggregate angle-associated block-accuracy contrast within each representation, computed using the angle contrast defined in Eq. (19). The learned-embedding aggregate crosses the encoder's source/target boundary and is not a clean standalone angle-sensitivity estimate.

Representation	Angle-associated contrast	95% interval
RAW	−0.2619	[−0.3587, −0.1619]
First difference	−0.1254	[−0.1952, −0.0540]
Learned embedding	−0.3778	[−0.4381, −0.3159]

Table 15 shows an adverse angle-associated contrast in the RAW spectrum, supporting a pre-encoder signal-level contribution. First differencing reduces the nominal magnitude of this contrast but does not establish invariant transfer.

The RAW contrast of −0.2619 should not be read as a second estimate of the −0.1111 factorial main effect in Section 4.3. The former is representation- and probe-specific, whereas the latter is the primary synchronised CNN factorial estimand. Their common contribution is directional: both show poorer performance at 45°, while their magnitudes are not directly comparable.

The learned-embedding contrast is larger numerically, but it is not a clean standalone estimate of angle sensitivity because it crosses the encoder's source/target boundary. Its magnitude is dominated by the sharp P3-to-P4 change and therefore mixes acquisition sensitivity with representation-transfer failure.

The evidence consequently supports a mixed localisation. Geometry-dependent signal degradation is present before learned encoding, while the source-trained representation fails to preserve some of the discrimination that remains linearly accessible at P4.

4.4.4 Acquisition position is strongly encoded in the source representation

A complementary source-only diagnostic shows that acquisition position is strongly linearly decodable from the frozen source representation: a linear position probe on P1–P3 embeddings reaches approximately 0.922 Accuracy, compared with a three-position uniform reference of 1/3 (Figure 12).

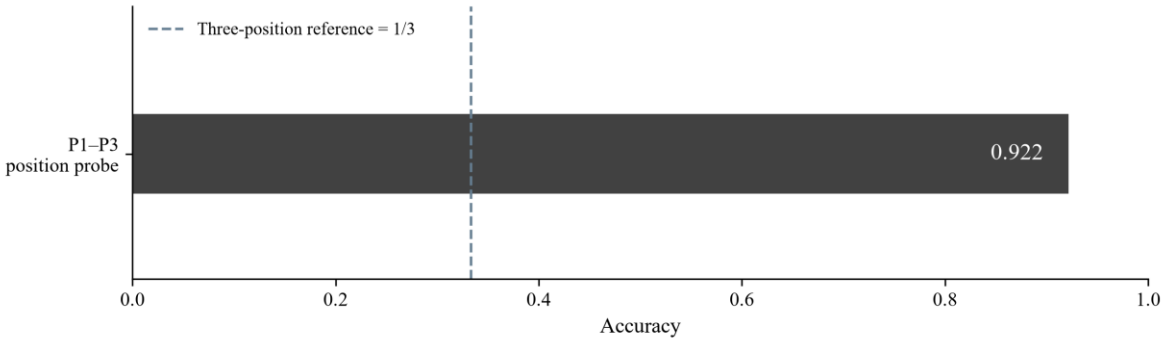

Figure 12. Source-position decoding from the frozen P1–P3 representation. The recorded mean linear-probe Accuracy is approximately 0.922; the dashed reference is 1/3 for three positions. This association does not establish a causal explanation of P4 failure.

Taken together with the P4 probe inversion, strong position decodability is consistent with geometry-entangled representation learning: the source embedding is highly TagID-discriminative within familiar geometries while also encoding acquisition position, yet TagID discrimination becomes substantially less linearly accessible at the unseen joint geometry.

This remains an association rather than a causal diagnosis. The DANN intervention in Chapter 5 was designed to suppress position information, but it did not reliably suppress the position probe or improve P4 performance. The causal role of position encoding therefore remains unresolved.

4.4.5 Supporting frequency-domain evidence

The probe analysis localises a representation-dependent component of the failure but does not identify which spectral structures are most sensitive to geometry. A complementary frequency-domain analysis was therefore used to provide physical context.

This evidence is deliberately secondary. The supplied 281-point representation is only conditionally mapped to approximately 5–8 GHz because the original frequency vector and exact crop indices are unavailable. More importantly, the pre-specified four-peak descriptor had a global missing-frequency rate of approximately 11.7%, exceeding the registered 5% extraction-validity threshold. It is therefore not used as a quantitative representation benchmark or causal mechanism test.

The surviving peak evidence is consistent with geometry-dependent changes in resonance visibility and apparent spectral structure, but it does not explain the P4 transfer collapse. Peak missingness is lowest at P4, so missing peaks cannot directly account for the held-geometry failure. Detailed peak-window definitions, the conditional GHz mapping and missingness breakdowns are retained as supporting material with Table 35 and Figure 18 in Appendix G.

4.4.6 Mechanism-driven negative follow-ups

Two additional studies tested specific physical explanations. Their conclusions remain here because they directly bound the mechanism claims; detailed execution records are deferred to the appendices.

Synthetic noise-floor augmentation. Source-side noise-floor augmentation tests whether the particular synthetic perturbation used here improves robustness. Its relevance to the proposed physical mechanism depends on how closely that perturbation represents the measurement disturbance affecting resonance visibility. The tested augmentation produced only small point gains relative to its control, with Accuracy 0.1666 ± 0.0264 and Macro-F1 0.1294 ± 0.0345. The gains were smaller than seed-to-seed variability, and the branch contained a disclosed protocol amendment retained in its frozen records. The result remains unconfirmed: it does not establish a reliable augmentation remedy or disprove a noise-related physical mechanism.

Peak-collision criterion. A second hypothesis was that geometry causes characteristic tag peaks to move close enough that specific TagID pairs become confused. The criterion predicted a collision for all 21 possible unordered TagID pairs. It therefore captured every pair belonging to the observed collapse sets, but also every pair in their non-collapse complements, yielding a 100% false-alarm rate under both reported collapse definitions. It therefore had no discriminative ability to explain which class pairs collapsed.

These negative outcomes bound the interpretation. The chapter can localise a joint acquisition- and representation-dependent problem, but it cannot reduce the failure to one validated RF mechanism. The detailed 13/252 reliable-separation statistic and tag-pair × peak × position breakdown are therefore supporting material rather than main-text evidence.

4.4.7 Chapter synthesis and transition to remedies

The historical results first establish that evaluation regime materially changes the apparent difficulty of the task, but they are not treated as the final transfer estimate. Under the governed source-only protocol, the selected C1 pipeline reaches only 0.1566 Accuracy and 0.1122 Macro-F1 at P4 and exhibits severe class collapse. Source-score reuse regret provides descriptive support for C1 within the recorded four-candidate score set; it does not establish target-optimal selection or quantify the failure avoidable by another source-only strategy.

The subsequent acquisition diagnostics refine how the failure should be interpreted. P4 is not uniquely difficult under matched within-position evaluation; P2 is comparably weak. The synchronised factorial analysis supports a reliable adverse association with 45°

acquisition while leaving the overall distance main effect and the angle-by-distance interaction unresolved. P4 is therefore scientifically distinctive as the unobserved 150-mm/45° joint geometry, not as a uniquely unreadable position.

Finally, the failure is not confined to one layer of the pipeline. Angle-associated degradation is already detectable before learned encoding, while the source-trained C1 representation fails at P4 to preserve some TagID discrimination that remains linearly accessible in RAW. Acquisition position is strongly decodable from the source embedding, although its causal role remains unresolved, and the tested peak-collision rule does not explain the observed class-collapse pattern.

Having used targeted follow-ups to bound specific physical explanations, Chapter 5 broadens the intervention question to standard source-only DG objectives and controlled target access.

5 Testing Possible Remedies

Chapter 4 established the governed P1–P3 → P4 transfer failure and localised it across the measured signal and the learned representation. This chapter changes from diagnosis to intervention. The experiments are organised as an **information-access ladder** rather than as a common performance leaderboard. Section 5.1 retains the strict source-only boundary and asks whether standard domain-generalisation objectives improve held-P4 transfer. Section 5.2 deliberately authorises limited labelled P4 information and asks how much calibration, readout replacement and encoder adaptation recover. Section 5.3 then distinguishes matched target-validation selection, historical full-outcome hindsight and within-condition target fitting to examine why apparently large target scores can arise and what those scores are licensed to mean. Section 5.4 integrates the resulting boundary map and transitions to the engineering implications developed in Chapter 6.

The ordering matters. Increasing target access can make a problem easier, but it also changes the scientific question. A higher score obtained after using P4 labels is therefore not an improved Strict-DG result; it is evidence from a different access class as defined in Section 3.2.

5.1 Can Source-Only Training Reduce Geometry Dependence?

The first intervention stage keeps the target boundary unchanged. P4 remains unavailable to fitting and model selection, and each method is judged only against an empirical-risk model trained under the same branch-specific conditions. The purpose is not to identify the numerically highest absolute P4 score. It is to ask whether a particular source-only intervention improves its own matched control.

Three observations from Chapter 4 motivate the tested remedies. First, source embeddings contain strong position information, motivating invariance- and adversarial-learning objectives. Second, source performance is heterogeneous and P2 is a comparatively weak source geometry, motivating a worst-group objective. Third, the factorial corpus contains the same physical conditions across source positions, motivating a cross-position interpolation study. IRM, DANN and GroupDRO test the first two ideas; the Mixup branch tests whether the third can be executed without changing the frozen split design.

5.1.1 Matched comparisons rather than an absolute leaderboard

Each completed intervention is compared with a branch-local ERM control trained under the same sampler, schedule, aggregation and seed framework. The matched effect is therefore (Eq. (22))

$$\Delta_{\text{method}} = M_{\text{intervention}} - M_{\text{matched ERM}}, \tag{22}$$

where M is the branch-defined P4 endpoint. Because the branch protocols differ, the three ERM values are not repeated estimates of one common canonical model. They are independently trained controls for the corresponding intervention.

The canonical C1 result from Section 4.2 remains the primary Strict-DG anchor for the dissertation, but it is not used as the numerical control for these later branches. This distinction prevents small implementation differences from being mistaken for method effects.

5.1.2 Invariant risk minimisation

Invariant risk minimisation (IRM) asks whether a representation can support a predictor that behaves consistently across the source environments [25]. Here P1, P2 and P3 are the environments. The penalty strength and its training schedule were selected using source evidence only; detailed development-grid and implementation checks are retained in Appendix B.

IRM does not provide a reliable P4 improvement. Its matched ERM reaches Macro-F1 0.118709 and IRM reaches 0.099379. The paired effect is

$$-0.019330,$$

with seed-inclusive 95% interval

$$[-0.066939,\ +0.025387].$$

The mechanism diagnostics are also unconvincing. Environment-risk dispersion increases by +0.014487 with interval [+0.005188, +0.022851], while the change in position decodability is −0.025397 with interval [−0.092063, +0.012698]. The tested IRM branch is therefore a valid negative intervention: the intended source invariance is not established and the held-P4 endpoint does not improve reliably.

5.1.3 Domain-adversarial training

The strongest source-side mechanistic motivation concerns position information. Section 4.4.4 showed that acquisition position is strongly linearly decodable from the frozen source embedding. Domain-adversarial neural training (DANN) therefore provides the most direct

intervention in this chapter: it adds a position classifier through a gradient-reversal layer so that the encoder is trained to preserve TagID information while making reader position harder to decode [24].

The adversarial weight $\lambda = 0.10$ was selected from P1–P3 evidence. The matched ERM and DANN arms were then evaluated on P4 under the same branch-specific training and seed framework.

DANN has a positive point effect, but its interval crosses zero. Table 16 shows no reliable manipulation of the proposed mechanism.

Table 16. DANN mechanism diagnostics.

Diagnostic	Matched ERM	DANN	Delta	95% interval / status
Position-probe Macro-F1	0.912464	0.905657	−0.006807	[−0.019896, +0.006788]
TagID-probe Macro-F1	0.234431	0.226745	−0.007686	[−0.026223, +0.010967]

The position-probe change is small and uncertain, and familiar TagID retention is not improved. DANN therefore produces neither reliable position suppression nor a reliable P4 benefit. Strong position decodability remains an association and a plausible component of the shift, but this intervention did not manipulate it strongly enough to identify a causal position-to-failure pathway.

5.1.4 Group distributionally robust optimisation

GroupDRO shifts optimisation toward the currently worst-performing source environment [26]. The motivation is source heterogeneity rather than a claim that one specific physical mechanism has already been established. P2 is a comparatively weak source geometry under the matched within-position analysis and, like P4, is acquired at 45°. A worst-group objective therefore tests whether protecting the disadvantaged source environment improves robustness at the held geometry.

The matched ERM reaches P4 Macro-F1 0.112439 and GroupDRO reaches 0.113844. The paired target effect is

$$+0.001405$$

with seed-inclusive 95% interval

$$[-0.042978,\ +0.039489].$$

Target block Accuracy changes by 0.000000 [−0.028571, +0.028571]. The source worst-position point effect is +0.024357 [−0.011581, +0.054235], but this does not translate into a reliable held-target gain. GroupDRO therefore does not provide evidence that emphasising the weakest source geometry repairs the P4 failure under this protocol.

5.1.5 Can cross-position interpolation help?

The final source-only hypothesis is that interpolation between matched physical conditions recorded at different source positions could expose the model to intermediate source geometries [27]. Because the raw acquisition is factorial, each source position contains the same 63 TagID–permittivity–surface conditions. The difficulty appears only after the frozen training splits are imposed.

The pre-specified feasibility criterion required more than 99% of possible cross-position parent pairs to remain available within each training split. Every independently constructed split provided only 2,100 of 4,200 opportunities, exactly 50%. The raw data are complete; the incompatibility is split-induced because independently assigned within-position training subsets share only 21 of 42 eligible condition keys.

No Mixup performance run was therefore conducted. Changing the split after observing the failed gate would have changed the proposed experiment. This result is evidence of **protocol infeasibility under the stated pairing rule**, not evidence that Mixup itself is ineffective. A future jointly matched split could test the method as a new experiment.

5.1.6 What do the source-only interventions establish?

Table 17 shows no reliable matched improvement from the three executed source-only interventions, while the cross-position Mixup branch produces no performance estimate. The appropriate conclusion is deliberately narrower than saying that domain generalisation fails in general.

Table 17. Source-only intervention effects against branch-matched ERM controls. Mixup contributes feasibility evidence only.

Intervention	Matched ERM P4 Macro-F1	Intervention result	Matched effect [95% CI]
IRM	0.118709	0.099379	−0.019330 [−0.066939, +0.025387]
DANN	0.101116	0.117153	+0.016037 [−0.035578, +0.068624]
GroupDRO	0.112439	0.113844	+0.001405 [−0.042978, +0.039489]
Condition-matched cross-position Mixup	—	2,100 / 4,200 eligible parent pairs (50%)	No performance result; >99% feasibility gate not met

The result has two layers of interpretation. First, it is consistent with the broader domain-generalisation literature: under strict source-only model selection, specialised DG objectives do not automatically outperform a well-controlled ERM reference [6]. Second, the present sensing problem contains an additional physical complication. Section 2.2 formalised the retained measurement as (Eq. (23))

$$r = \mathcal{M}(y, e, s, d, \varepsilon), \qquad x = T(r), \tag{23}$$

so reader geometry d enters the physical measurement process before the learned transformation T. Geometry-associated variation can therefore reshape spectral structure that also carries TagID information rather than acting only as a separable nuisance appended after measurement. Chapter 4 showed that performance degradation is already detectable before learned encoding and that the source-trained representation does not preserve all of the TagID discrimination that remains linearly accessible at P4.

These observations make simple invariance or alignment non-trivial, but they do not prove that physical entanglement is the unique cause of the negative DG results. In particular, DANN did not reliably suppress position information, so the causal role of position encoding remains unresolved. The supported conclusion is that **the tested source-only objectives did not reliably repair this held-geometry failure under the available three-source design**.

An additional source-only follow-up tested whether self-supervised pre-training with permitted unlabelled external chipless-RFID data could improve P4 transfer. None of the tested SSL treatments provided a reliable or practically meaningful benefit over its branch-specific supervised control. Because the external corpus uses a different tag codebook, the complete treatment comparison and its portability boundary are retained in Appendix E.3.

5.2 How Much Does Limited Target Calibration Help?

Section 5.1 kept P4 outside development. This section deliberately changes the access class by authorising labelled target support. The engineering question is now different: if a reader can be calibrated after installation, how much labelled P4 information is required before held-condition recognition improves?

These results are target-assisted diagnostics. They quantify what controlled target access buys and cannot be substituted for the Strict-DG result of Section 4.2.

5.2.1 Few-shot calibration and the support/query boundary

The historical few-shot study uses the frozen 256-dimensional penultimate representation of the source-only C1 encoder. Encoder weights, first-difference preprocessing and the source model are fixed; only the declared adapter sees labelled P4 support.

The nine P4 permittivity–surface combinations are arranged into 18 pre-specified episodes. In each episode, five combinations form the support pool and the remaining four form the query set. The query set therefore contains 28 complete physical condition blocks, or 1,400 rows. Support sizes are 0, 7, 21 and 35 labelled measurements for 0-, 1-, 3- and 5-shot respectively.

The leakage boundary is stricter than row disjointness. Although only selected measurements become labelled support, **every complete physical condition block from the support pool is excluded from query evaluation**, including blocks from which no support measurement was selected. Query performance therefore cannot benefit from other repeat sweeps of the same physical condition. Isolation was additionally checked using sample identity, condition block, repeat group, exact signal digest and source-row identity; the 54 recorded isolation checks passed.

Two low-capacity adapters are compared. The target-only prototype classifier averages support embeddings by class and classifies queries by squared Euclidean distance. The source-anchored linear head updates only the final linear rule under a penalty toward the frozen source solution.

5.2.2 Historical 1/3/5-shot result

Table 18. Historical few-shot target calibration. Parenthesised values are population standard deviations over the 18 episode means from one reused P4 corpus.

Method	Shot	Labels	Accuracy	Macro-F1
Frozen source head	0	0	0.156635 (0.023472)	0.107257 (0.025602)
Target-only prototypes	1	7	0.142698 (0.037640)	0.112159 (0.033743)
Target-only prototypes	3	21	0.165429 (0.021727)	0.145141 (0.023730)
Target-only prototypes	5	35	0.162635 (0.029901)	0.144360 (0.029465)
Source-anchored head	1	7	0.147476 (0.031605)	0.120852 (0.027799)
Source-anchored head	3	21	0.149754 (0.024019)	0.129295 (0.021783)
Source-anchored head	5	35	0.157651 (0.040010)	0.138984 (0.036008)

The zero-shot Macro-F1 in Table 18 is not expected to equal the full-P4 canonical Macro-F1 of Section 4.2. The few-shot study evaluates episode-specific query partitions and aggregates over those episodes, whereas the canonical Strict-DG result is computed on the full held-P4 evaluation. The two values therefore have different aggregation paths even though both use the frozen source model.

Table 19 gives the more informative paired-episode comparison, also shown alongside the adaptation curves in Figure 13.

Table 19. Paired episode changes in Macro-F1 over the same 18 episodes.

Comparison	Mean delta	95% interval	Positive / negative episodes
Prototypes, 0 → 1 shot	+0.004902	[−0.012224, +0.020574]	11 / 7
Prototypes, 0 → 3 shots	+0.037884	[+0.022224, +0.056097]	17 / 1
Prototypes, 0 → 5 shots	+0.037103	[+0.019806, +0.054516]	16 / 2

Comparison	Mean delta	95% interval	Positive / negative episodes
Head, 0 → 1 shot	+0.013594	[+0.000978, +0.025817]	12 / 6
Head, 0 → 3 shots	+0.022038	[+0.011519, +0.032641]	16 / 2
Head, 0 → 5 shots	+0.031727	[+0.012092, +0.050825]	14 / 4
Head − prototypes at 3 shots	−0.015846	[−0.031515, −0.000718]	6 / 12

Three labelled measurements per class increase prototype Macro-F1 from approximately 0.1073 to 0.1451 within the episode analysis. The point gain is real within that reused corpus, but its absolute scale remains small and Accuracy remains close to the balanced seven-class reference. The bootstrap intervals resample 18 episodes drawn from one P4 corpus; they are not intervals over independent acquisition campaigns.

At five shots, the mean support cross-entropy becomes very small while the adapted head moves substantially from the source solution, and some units improve support fit while query Macro-F1 falls. This pattern is consistent with support-set overfitting, although it does not establish overfitting as the unique mechanism.

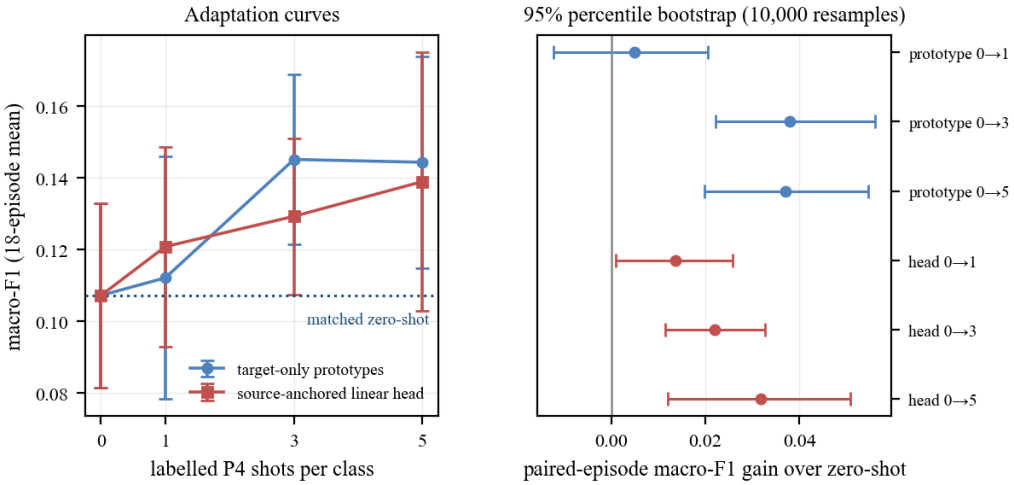

Figure 13. Historical few-shot calibration against the episode-specific zero-shot reference. The paired episode comparison is the inferential view; all episodes reuse one P4 corpus. Left-panel error bars show the population SD (ddof = 0) over the 18 episode-level means, each itself averaged over five seeds; right-panel intervals are 95% percentile intervals from 10,000 paired-episode bootstrap resamples of the same 18 episodes.

5.2.3 Does deliberate coverage of P4 nuisance conditions help?

The historical few-shot analysis answers a label-dose question. A separate follow-up asks a different question: at the same label budget, is it useful to choose support so that P4's within-target permittivity and surface conditions are deliberately covered?

This motivation is distinct from the angle–distance analysis of Section 4.3. P4 has fixed angle and distance, but it still contains nine permittivity–surface combinations. The follow-

up therefore uses Table 20 to compare random support with joint permittivity/surface coverage at matched 3-shot budget under a condition-block-disjoint protocol.

Table 20. Factor-aware versus random support at matched 3-shot budget.

Support-selection strategy	Block Macro-F1	Paired effect [95% CI]
Random support	0.129095	Reference
Joint permittivity/surface coverage	0.124103	−0.004992 [−0.039200, +0.029010]

The tested coverage strategy provides no reliable advantage over random support. The earlier historical 3-shot value of 0.145141 and the later random-support value of 0.129095 should not be compared directly: the studies use different partition constructions and aggregation rules. The former addresses label dose; the latter is the matched test of support-selection strategy.

5.2.4 Do larger calibration budgets produce a reliable dose response?

A later target-label ladder increases the support budget through 0, 7, 10, 20, 21, 35, 50, 100, 200 and 500 labels using nested support prefixes and a fixed cosine-prototype adapter.

Three fixed Latin-square query folds each withhold 21 of the 63 physical condition blocks; the other 42 blocks supply support. Support and query conditions, rows and exact signal hashes are disjoint within a fold. Budgets 7, 10, 20, 21 and 35 cover that many support blocks; budgets from 50 onwards cover all 42 available blocks, with additional labels adding repeated sweeps rather than new conditions. Each of the 3,150 rows is queried out of fold. Source preprocessing and the encoder remain frozen; query labels are used only for final scoring, with no target validation search.

Row metrics are computed from the pooled out-of-fold predictions for each source-seed/support-selection pair, then averaged over five source seeds and 20 support selections. Uncertainty uses 10,000 paired, TagID-stratified condition-block bootstrap replicates with crossed source/support-seed resampling (Appendix B.5).

Prototype calibration is non-monotonic. Macro-F1 peaks at approximately 0.135560 with 20 labels and falls to approximately 0.102347 with 500 labels.

More target labels therefore do not produce a simple monotonic recovery under the tested prototype protocol. The result does not establish a smooth extension of the historical 3-shot improvement: the adapter and query design differ. Below full block coverage, label count and condition coverage also vary together. It does not show that more target data could never help under another calibration design.

5.2.5 Is the transferred readout the main bottleneck?

Chapter 4 showed that, at P4, the raw spectrum retains more linearly accessible TagID discrimination than the source-trained C1 embedding. A remaining question is whether the frozen representation is nevertheless adequate once the source-trained decision rule is replaced by a target-trained one.

At 500 labelled P4 samples, a target-trained linear softmax readout on the frozen 256-dimensional C1 representation reaches Macro-F1 0.142740. Relative to the 500-label prototype result, the point effect is +0.040393, but the 95% interval [−0.040925, +0.116499] crosses zero.

The result is consistent with readout mismatch as a possible contributor, but it does not establish a reliable readout effect. A target-trained decision boundary gives only modest and uncertain point recovery and does not produce a reliable held-condition rescue.

5.2.6 Does updating the encoder help?

The final target-assisted intervention relaxes the constraint further by allowing the representation itself to adapt. Table 21 summarises the results at 500 labelled target samples:

Table 21. Readout and encoder adaptation at the 500-label target budget.

Adaptation level	P4 held-condition Macro-F1	Interpretation
Frozen encoder + target-trained linear readout	0.142740	Modest uncertain point recovery
Partial encoder fine-tuning	0.131861	No improvement over frozen linear readout
Full encoder fine-tuning	0.129613	No improvement over frozen linear readout

Partial and full fine-tuning have negative point effects relative to the frozen linear readout, and their uncertainty intervals cross zero. Under the tested support budget, allowing the encoder itself to change therefore does not rescue held-condition performance.

Readout mismatch may contribute. The tested readout replacements provide only limited recovery, and updating the encoder under the tested calibration budget does not provide a reliable rescue. These procedures do not identify the maximum TagID information recoverable by every possible decoder.

5.2.7 What does limited target access establish?

The target-label studies answer related but distinct questions. Historical 1/3/5-shot episodes show a small episode-level gain at 3 and 5 shots. Deliberate permittivity/surface coverage does not outperform random support at matched budget. Larger prototype budgets do not

yield a monotonic calibration curve. A trainable linear readout gives modest uncertain point recovery, while partial and full encoder fine-tuning do not establish a reliable improvement in the endpoint.

The combined conclusion is therefore not that target labels are useless. It is that **controlled P4 labels produce limited, unstable and protocol-sensitive recovery rather than a simple rescue**. Readout mismatch may contribute, but the tested procedures do not establish the best recovery attainable by other decoders or calibration designs.

5.3 What Can Hindsight Reveal—and What Can It Not?

This section examines three distinct information-access arrangements. The matched A0/A1 study authorises P4 validation labels for selection while protecting P4-Test evaluation labels. The historical retrospective analysis uses full P4 outcomes to promote a readout after the fact. A2 additionally fits readouts on labelled P4-Adapt data and evaluates within represented physical conditions. These designs answer different questions and are not prospective Strict-DG results.

The central rule of this section is therefore **within-lineage interpretation**. Cross-lineage score differences are decomposed rather than treated as before/after improvements.

5.3.1 Matched target-informed selection within one development path

A fairness-controlled follow-up isolates the effect of permitting P4 validation labels to influence selection while holding the development path fixed. Twenty-five paired units combine source seeds 42–46 with target-split seeds 11, 23, 37, 53 and 71. These splits reuse one corpus and are not independent target acquisitions.

A0 fits and selects using source evidence only. A1 uses P4-Val labels solely to choose among the same frozen source-fitted learned-head, cosine-NCM and Euclidean-NCM candidates; it does not fit a readout or update the encoder. The full selection registry is frozen and hashed before P4-Test labels are scored. All 63 physical conditions occur in P4-Adapt, P4-Val and P4-Test for all arms, although exact signal-hash groups do not cross roles. Table 22 reports this controlled target-assisted selection diagnostic within represented conditions, not full-P4 retrospective selection or held-condition transfer.

Table 22. Matched source-only and target-informed selection within one development path.

Arm	Macro-F1	Matched interpretation
A0 source-only selection	0.111437	Matched diagnostic control
A1 target-informed selection	0.113742	A1 − A0 = +0.002304 [−0.001057, +0.006679]

The matched difference is small and its interval crosses zero. Within this development path, target-informed selection does not provide a reliable improvement. It estimates the effect of the authorised validation-based selection within this fixed development path and within-condition partition protocol.

5.3.2 Historical retrospective carrier and access boundary

A separate historical analysis produced a much larger target score, but it belongs to a different representation lineage. Its purpose is forensic: to show why a cross-lineage headline contrast cannot be assigned to target access alone.

Table 23 summarises the access boundary for a frozen historical representation carrier produced by an earlier line of work. The 7,350 source embeddings are row-normalised, grouped by TagID, averaged within class and normalised again to form a class prototype. Each of the 3,150 P4 embeddings is classified by maximum cosine similarity to those **source-derived** prototypes. There are no target-derived prototypes, no trainable parameters and no optimiser steps in this readout. What was selected retrospectively was the readout formulation itself from competing normalisation and whitening variants.

Table 23. P4 information access in the historical retrospective analysis.

Stage	Access
Source prototype construction	Source only within the historical carrier lineage
Prediction inference	Target features, unlabelled
Readout promotion / selection	Retrospective; full P4 outcomes
Final metric reporting	Target labels

The same 3,150 P4 examples that influenced readout promotion also supply the reported target metrics. The result is therefore in-sample with respect to model/readout selection.

The retained retrospective readout reaches Accuracy 0.577143 and Macro-F1 0.583569 for one seed. It has zero zero-recall classes. This number is deliberately **not** promoted into the Strict-DG benchmark: the representation lineage differs, full P4 outcomes influenced readout promotion, and there is no independent target test set or multi-seed uncertainty estimate.

5.3.3 Why the retrospective score is not a target-access gain

Table 24 decomposes the visually large difference between canonical Strict-DG Accuracy 0.156635 and retrospective Accuracy 0.577143; it is not a matched target-access effect. The same historical carrier already reached approximately 0.535556 Accuracy under its recorded source-only selector.

Table 24. Cross-lineage decomposition of the retrospective headline gap.

Configuration	Accuracy	Macro-F1	Target information in selection
Canonical Strict-DG C1, five-seed mean	0.156635	0.112231	None
Source-only selector on historical carrier	0.535556	0.538281	None in readout selection; carrier development lineage differs
Retained retrospective readout	0.577143	0.583569	Full P4 outcomes

Most of the recorded numerical difference from canonical C1 is already present in the historical carrier before retrospective readout promotion. The headline gap therefore cannot be interpreted as the effect of target-informed readout selection alone. Because representation and other development details differ between lineages, their individual contributions to that gap are not identified.

One provenance caveat remains. An independent reimplementation of the historical source-only selector on the same carrier produced 0.4921 rather than the recorded 0.5356. The exact intermediate value is therefore only approximately corroborated. The qualitative conclusion is unchanged: most of the apparent gap precedes the target-informed readout choice.

The historical carrier does not demonstrate that canonical C1 contains substantial recoverable P4 information, and it does not contradict the encoder-clean representation diagnostic in Section 4.4. The canonical-lineage readout and fine-tuning interventions in Section 5.2 are the relevant tests of that question.

5.3.4 Within-condition interpolation: how high can a permissive target score become?

The third target-informed arm deliberately permits calibration and evaluation to reuse the same physical condition blocks. Table 25 reports this arm to show how misleading a high target score can become when the evaluation question changes.

Table 25. Within-condition target calibration and held-condition controls.

Comparison	Macro-F1	Interpretive licence
Target-fitted prototypes (A2)	0.402750	Within-condition interpolation
Target-fitted linear rule (A2)	0.983494	Dependence-inflated; not transfer
Leave-permittivity-out control	0.1432	Held-condition diagnostic
Leave-cell-out control	0.0865	Held-cell diagnostic

All 63 physical conditions occur across the A2 adaptation, validation and evaluation partitions, creating strong exact or near-duplicate dependence. Under that permissive boundary, the linear rule reaches Macro-F1 0.983494 and the prototype rule reaches 0.402750; raw-signal and random-encoder controls can also become very strong.

Once condition dependence is removed, performance collapses toward the balanced seven-class reference. The result is therefore not evidence of transferable P4 recognition. It is a direct demonstration that spectacular interpolation accuracy can coexist with poor held-condition generalisation.

5.3.5 What does hindsight establish?

The three hindsight analyses resolve different ambiguities.

First, matched P4-validation selection has a small uncertain effect within one development path. Second, most of the historical numerical gap is already present before retrospective readout promotion; the contributions of representation and other development differences are not identified. Third, near-perfect performance occurs when represented physical conditions reappear, whereas the recorded leave-condition controls collapse.

Hindsight therefore does not provide a replacement benchmark. Its value is diagnostic: it shows how **access, lineage and dependence structure can change the meaning of a target score more strongly than the score itself reveals**.

The retrospective evidence is bounded by one reused target corpus, one historical seed, incomplete carrier provenance and no independent target test set. Those limitations affect the strength of the historical number, but they do not alter the matched A0/A1 conclusion or the central warning against cross-lineage and within-condition leaderboard comparisons.

5.4 Chapter Synthesis and Transition

Chapter 5 has increased information access step by step rather than searching for one best score. Read in that order, the negative results form a boundary map rather than a sequence of failed experiments.

Table 26. Intervention boundary map across increasing information access.

Information access	Intervention / analysis	Main result	Supported interpretation
Strict source-only	IRM	Matched effect −0.0193; interval crosses zero	No reliable held-P4 benefit
Strict source-only	DANN	Matched effect +0.0160; no reliable position suppression	No reliable benefit; causal role of position remains unresolved
Strict source-only	GroupDRO	Matched effect +0.0014; interval crosses zero	Protecting the weakest source group did not rescue P4
Strict source-only	Condition-matched cross-position Mixup	50% parent coverage; feasibility gate failed	No performance conclusion
Limited labelled P4	Few-shot prototypes / linear head	Small episode-level gains	Limited calibration benefit, not a rescue

Information access	Intervention / analysis	Main result	Supported interpretation
Limited labelled P4	Factor-aware support	−0.0050 versus random; interval crosses zero	No reliable advantage for the tested coverage strategy
Larger labelled P4	Prototype calibration	Non-monotonic through 500 labels	No stable label-dose law
Larger labelled P4	Target-trained readout / encoder fine-tuning	Linear readout modest and uncertain; fine-tuning no better	Readout mismatch may contribute; the tested procedures do not establish a reliable rescue
P4-validation selection	Matched A0/A1; frozen source-fitted readouts	+0.0023; interval crosses zero	Small uncertain selection effect; protected P4-Test scoring within represented conditions
Retrospective historical carrier	Source-only → retrospectively promoted readout	0.5356 → 0.5771 Accuracy	Most of the numerical gap precedes promotion; individual development contributions are not identified
Within-condition target access	A2 linear rule	Macro-F1 0.9835	Interpolation among represented physical conditions; not transfer

The source-only interventions do not support a reliable algorithmic rescue under the present three-source design. Limited target labels provide small, unstable or protocol-sensitive recovery. Replacing the readout does not remove the failure, and updating the encoder under the tested budget does not reliably improve held-condition performance. Strong retrospective or within-condition scores are achievable only after the access boundary or development lineage changes, and those numbers do not answer the original Strict-DG question.

Taken together with Chapter 4, these results argue against treating the P4 failure as a classifier-only defect. Reader geometry enters the physical measurement before learned representation; the source-trained embedding does not preserve all of the discrimination that remains accessible at P4; and the tested readout and encoder interventions do not provide a reliable repair. The engineering emphasis therefore shifts toward **joint measurement, representation and evaluation design**, rather than loss-function substitution alone.

Two supporting studies test the boundary of this conclusion without changing it. Appendix E reports the independent four-class external corpus and self-supervised follow-up; these establish scoped pipeline portability but not transfer of the seven-class model. Appendix F reports the OpenEMS design exploration; its confirmation criteria were not met and the simulation omits the complete measurement chain, so it is retained as exploratory design evidence rather than as a demonstrated hardware rescue.

Chapter 6 integrates these bounded intervention results with the acquisition and representation diagnostics, states the remaining limitations, and identifies the measurement and system-design changes required for a stronger future test.

6 Discussion, Limitations and Conclusions

This chapter brings the experimental programme back to the engineering question posed in Chapter 1. Chapters 4 and 5 established the held-geometry failure, localised where it appears, and tested several possible remedies under explicitly different information-access regimes. The purpose here is not to replay those results as a second Results section. Instead, the evidence is integrated into three questions: what the apparently contradictory performance figures mean when read under their correct evaluation regimes; what the combined signal, representation and readout diagnostics support about the failure; and what follows for future chipless-RFID measurement and system design.

The chapter then consolidates the principal limitations at the level at which they restrict the conclusions, before answering RQ1–RQ5 directly and identifying the next experiments that would most materially strengthen the evidence.

6.1 Integrated Discussion

6.1.1 From Familiar-Condition Learnability to Held-Geometry Failure

The central methodological result of the dissertation is that a performance figure cannot be interpreted independently of the split unit and information-access regime that produced it. Three headline results make this point particularly clearly.

The historical condition-grouped evaluation reached an Accuracy of 0.7344 under a regime in which training and test conditions came from the same broad measurement population. Under the governed P1–P3 → P4 Strict-DG protocol, the selected C1 model instead reached Accuracy 0.1566 and Macro-F1 0.1122, with severe class collapse. Later, a deliberately within-condition target-assisted diagnostic reached Macro-F1 0.9835 because the target partitions shared physical conditions already represented elsewhere in the target data. These numbers are not competing estimates of one model's quality. They answer different questions.

The 0.7344 result establishes familiar-condition learnability. The 0.1566/0.1122 result establishes the difficulty of transfer to the held joint 150-mm/45° geometry when P4 is excluded from source-only development. The 0.9835 result demonstrates how high a score can become when the evaluation permits interpolation among already represented physical conditions. Read together, they show why repeated sweeps cannot be treated as independent replication and why random or familiar-condition accuracy cannot be used as a deployment estimate for an unseen installation.

The same distinction explains the retrospective and target-assisted results in Chapter 5. Additional P4 information can change the model, the representation, the readout or the condition overlap available during evaluation. Once that happens, the scientific question has changed. The central lesson is therefore not that one score is more "correct" than another, but that each score is valid only within the access and dependence structure under which it was produced.

6.1.2 A Joint Signal–Representation–Readout Limitation

The mechanism evidence is most coherent when organised across the three levels summarised in Table 27. Together they constrain interpretation without identifying a unique mechanism.

At the **measured-signal level**, the synchronised acquisition-factor analysis supports a reliable adverse association with 45° acquisition, and the RAW representation also degrades at the two 45° positions. This supports an acquisition-dependent signal contribution before learned encoding. It does not establish angle as the dominant physical factor, exclude a role for distance, or isolate one electromagnetic mechanism.

At the **representation level**, the encoder-clean P4 comparison provides the clearest result. Under the same held-condition linear-probe protocol, RAW reaches Macro-F1 0.2427 whereas the source-trained C1 embedding reaches 0.1472; the paired RAW-minus-embedding difference is +0.0951 with a 95% interval of [0.0173, 0.1747]. The supported interpretation is that C1 does not preserve at P4 all of the TagID discrimination that remains linearly accessible in the raw spectrum. This is a representation-transfer limitation, not proof that TagID information has been destroyed or is absent from the embedding.

The source embedding also carries strong acquisition-position information: a source-only position probe reaches approximately 0.922 Accuracy on P1–P3. Taken together with the P4 probe result, this is consistent with geometry-entangled representation learning. It remains associative. DANN did not reliably reduce position decodability, so the dissertation does not establish that position encoding itself causes the held-P4 failure.

At the readout level, controlled P4 access provides only modest evidence for decision-boundary mismatch (Table 27). Recovery remains uncertain and partial or full encoder fine-tuning does not reliably improve held-condition performance. Readout mismatch may therefore contribute, but the tested procedures do not identify the maximum information recoverable by every possible decoder.

Table 27. Three-level failure synthesis integrating acquisition and representation diagnostics from Chapter 4 with readout and adaptation evidence from Chapter 5.

Level	Evidence	Supported conclusion
Measured signal	Reliable adverse angle association; RAW already degrades at 45° (Sections 4.3–4.4).	An acquisition-dependent signal contribution is supported; no unique RF mechanism or complete physical loss of TagID discrimination is established.
Representation	At encoder-clean P4, RAW Macro-F1 = 0.2427 and C1 embedding Macro-F1 = 0.1472; paired difference +0.0951 [0.0173, 0.1747] (Section 4.4.2).	C1 preserves less linearly accessible TagID discrimination at P4; linear-probe failure does not establish that all TagID information is absent.
Readout	At 500 target labels, linear-readout Macro-F1 = 0.142740; recovery is uncertain, and partial/full fine-tuning gives no reliable rescue (Sections 5.2.5–5.2.6).	Readout mismatch may contribute; the tested procedures do not establish a reliable rescue or the best recovery attainable by every decoder.

Table 27 is therefore not a claim that three independent failures have each been causally proven. It is a localisation of where the empirical evidence places limitations in the sensing-and-learning chain. The strongest supported interpretation is a joint acquisition-dependent signal and representation-transfer problem, with possible residual readout mismatch.

6.1.3 What the Negative and Unconfirmed Results Constrain

The negative, blocked and unconfirmed branches are useful because they narrow the space of explanations and remedies without being promoted beyond their evidence.

As summarised in Table 26, IRM, DANN and GroupDRO constrain the tested interventions rather than the method families in general: none provides a reliable held-P4 benefit under its branch-matched protocol. DANN additionally fails to produce reliable position suppression, so it does not resolve whether the strong source-position decodability has a causal role in the P4 failure.

The factor-aware support experiment found no reliable advantage over random support at the matched 3-shot budget. This constrains the tested joint ER/surface coverage rule; it does not show that target-support identity can never matter. The larger target-calibration curve is similarly important because its non-monotonic behaviour does not support a simple monotonic calibration law under this design; it does not rule out benefit from another target-sampling or calibration design.

The proposed cross-position Mixup experiment contributes a different kind of result. The raw acquisition contains the relevant factorial correspondence, but independently constructed training splits retained only 50% of the required parent pairs, below the pre-specified feasibility gate. No training result was therefore produced. The lesson is methodological: pairing-dependent methods require cross-position correspondence to be

declared and preserved before the split is frozen, rather than reconstructed after observing that the proposed pairing is infeasible.

A related protocol lesson emerged from the factorial audit. Exact paired inference required validation assignment to be independent of reader position; when an inherited assignment coupling was identified, the full 60-run factorial experiment was rerun with synchronised folds and seeds before inference. Together, these episodes show that data-governance choices are not administrative details: split construction and randomisation keys can determine whether a proposed scientific comparison is valid.

The broader programme also explored complementary directions beyond the primary source-only and target-assisted branches. Appendix E tests external-corpus pipeline portability and self-supervised pretraining, while Appendix F tests a simulation-led tag redesign. Neither supplies a validated general remedy for the seven-class held-geometry problem under its own evidence boundary. These studies extend the search space considered here, but they do not exhaust the possible algorithmic, sensing or hardware solutions.

6.1.4 Engineering Implications

The results support several practical implications, but their strength should match the underlying evidence.

Characterise geometry rather than distance alone. The present campaign supports a reliable adverse association with 45° acquisition, while the overall distance main effect is not confirmed and the negative distance simple effect at 0° remains supported for pooled condition-block Macro-F1 but not for Accuracy under seed-inclusive uncertainty. Installation characterisation should therefore record and control both angle and range rather than assuming that one scalar standoff distance is sufficient to describe deployment difficulty.

Do not size a system from familiar-condition accuracy. The difference between familiar-condition, Strict-DG and within-condition results shows that classification performance is inseparable from the evaluation regime. A deployment-oriented specification should keep repeated physical conditions together and include held geometries that did not participate in model development.

Treat target calibration as an intervention, not an assumed rescue. Limited target labels provide some descriptive improvements in particular episodes, but the larger calibration curve is non-monotonic and target-trained readout or encoder updating remains weak or uncertain. Site-specific calibration may be useful, but its benefit must be measured under held-condition queries rather than assumed from the number of labelled target examples.

Treat physical code separation as a design hypothesis requiring direct validation. The four-peak descriptor provides motivation to examine code separation, but it failed its registered extraction-validity gate. The 13/252 separation statistic therefore describes that tested descriptor rather than physical codebook margin itself and should not be used as a fabrication criterion.

For future data campaigns, physical condition blocks should remain the atomic split unit; independent campaigns should be distinguished from repeat sweeps; and any cross-position correspondence required by a pairing-based method should be declared before the split is frozen. These practices follow directly from the evaluation and feasibility findings of the present work.

6.2 Limitations

The limitations that most materially restrict the conclusions fall into three connected groups. Detailed run-level provenance and implementation records remain in Appendix D; the purpose here is to state how the limitations constrain scientific interpretation.

6.2.1 Experimental Scope and External Validity

The principal transfer claim is based on one depolarizing-tag family, one measurement campaign and one held target geometry. P1–P3 provide only three source domains, while the acquisition design samples two distance levels and two angle levels. The supported angle contrast therefore describes the difference between 0° and 45° in this campaign; it does not establish the shape, monotonicity or threshold of angle sensitivity outside those sampled levels. Likewise, P4 is one held joint 150-mm/45° geometry, not evidence that all reader-position shifts behave similarly.

The geometry evidence is observational within the recorded campaign. The factorial analysis provides a reliable adverse angle association, but the dissertation does not claim that angle is the unique or dominant electromagnetic cause. The overall distance main effect is not confirmed, the 0° simple distance effect remains negative for pooled condition-block Macro-F1 while its seed-inclusive Accuracy interval crosses zero, and the interaction remains unconfirmed and floor-limited.

These constraints limit external validity rather than the internal interpretation of the governed P1–P3 → P4 experiment. The most direct way to strengthen the claim would be a prospective acquisition containing at least one additional held geometry fixed before model development and target scoring, preferably with more than two sampled values of distance and angle.

6.2.2 Data and Statistical Limitations

The 12,600 measurement rows arise from 252 physical condition blocks with fifty repeated sweeps per block. Condition-block grouping prevents repeated sweeps of the same physical condition from crossing the principal train/test boundaries, but it does not create 252 independent acquisition campaigns. Repeated measurements characterise within-condition variability; they should not be counted as independent evidence about new installations.

The available processed representation is magnitude-only. Complex phase and richer polarimetric channels are not available for the principal analysis. In addition, the authoritative original frequency vector and crop indices for the supplied 281-point representation are unavailable, which is why the frequency mapping in the peak study is conditional. The pre-specified peak descriptor also failed its global extraction-validity gate, limiting the physical interpretation that can be drawn from that branch.

Uncertainty is study-specific. The primary Strict-DG result reports dispersion over five training seeds; the factorial analysis uses paired condition blocks with seed sensitivity; the historical few-shot intervals reuse episodes from one P4 corpus; source-selection regret has only three held-source events; and the retrospective historical carrier is a single lineage instance. None of these uncertainty summaries should be read as a confidence interval over independent devices, laboratories or future acquisition campaigns.

A stronger statistical basis therefore requires new independent acquisition rather than additional resampling of the existing P4 corpus. More repeated use of the same target data can refine within-corpus estimates, but it cannot substitute for external acquisition-level replication.

6.2.3 Protocol, Provenance and Interpretive Boundaries

The final Strict-DG implementation is source-only, but the wider research programme had previously examined P4 in other lineages. The result is therefore a governed source-only evaluation rather than a claim of a historically untouched target. The 60 source-selection units were replayed and independently re-scored rather than fully retrained, whereas the five final Strict-DG models were retrained. Some secondary branches also rely on reconstructed or replayed evidence; Appendix D records their exact status and provenance.

The intervention branches are not one common leaderboard. IRM, DANN and GroupDRO use their own branch-matched ERM controls, and target-assisted, retrospective and within-condition studies authorise different amounts of P4 information. Their results answer different questions and cannot be combined into a single ranking by absolute score.

The mechanistic interpretation remains non-causal. Position decodability is consistent with geometry-entangled representation learning, but the DANN intervention did not produce reliable position suppression, so causal responsibility remains unresolved. The OpenEMS study is likewise exploratory: its confirmation criteria were not met, the complete measurement chain is not modelled, and no measured hardware or classification improvement is claimed.

These boundaries define the level at which the conclusions are valid. A prospective untouched-target campaign, a successful representation intervention that demonstrably reduces position information while retaining source TagID discrimination, and measured validation of any physical redesign would each remove a different part of the present uncertainty.

6.3 Conclusions and Future Work

6.3.1 Answers to the Research Questions

The five answers below map directly to the research questions and decision evidence defined in Chapter 1.

RQ1 — Why do familiar-condition results differ from transfer to a held reader geometry?

The grouped familiar-condition evaluation tests unseen condition blocks drawn from reader geometries already represented during development, whereas Strict-DG withholds the complete P4 geometry. Within-condition interpolation additionally permits reuse of physical conditions and is a separate evaluation regime. High familiar-condition performance therefore does not establish held-geometry transfer.

RQ2 — Can a recipe developed on P1–P3 generalise to P4, and was its source-only selection defensible?

The source-only selector chose first-difference C1 from P1–P3 evidence. The source-score reuse regret analysis provides descriptive support for C1 within the recorded four-candidate score set, not independent validation of nested held-source selection. It does not establish P4-optimal selection or determine how much target failure another source-only strategy could avoid. The selected model nevertheless reached only Accuracy 0.1566 and Macro-F1 0.1122 at P4 over five seeds and exhibited severe class collapse.

RQ3 — Where is the transfer failure located?

The evidence identifies limitations at several tested stages without establishing a unique explanation. Acquisition-dependent degradation is already visible in the measured signal; at

P4 the source-trained embedding preserves less linearly accessible TagID discrimination than RAW; and target-trained readout changes provide only modest and uncertain recovery. Readout mismatch may contribute, but the tested procedures do not bound every possible decoder. Position is strongly decodable from the source embedding, but its causal role remains unresolved.

RQ4 — Can standard DG objectives or controlled target information reliably repair the failure?

Not under the tested protocols. IRM, DANN and GroupDRO provide no reliable branch-matched improvement, while the proposed Mixup experiment fails its pre-specified pairing feasibility gate and therefore yields no performance conclusion. Few-shot calibration, larger target-label budgets, trainable readouts and encoder fine-tuning provide only limited, unstable or uncertain held-condition recovery.

RQ5 — What do the findings imply for future chipless-RFID measurement and system design?

Robust identification should be treated as a joint measurement, representation and evaluation problem. Future systems should characterise realistic reader geometry, preserve physical condition blocks in evaluation, and assess target calibration under held-condition queries rather than assuming that more target labels guarantee recovery. Physical code separation remains an additional design hypothesis that requires stronger direct validation.

6.3.2 Overall Conclusion and Claim Boundary

This dissertation makes four linked contributions: an access-aware, condition-block-grouped, strict source-only evaluation framework for reader-geometry shift; an empirical separation of familiar-condition performance from held-geometry transfer; a three-level localisation of the observed failure across measured signal, learned representation and readout; and a controlled evaluation of remedy and calibration boundaries under explicit information access. Together, these contributions shift the emphasis away from reporting a single best classification score and toward establishing what kind of generalisation is being measured, whether a source-selected model transfers to a held joint geometry, where transfer fails across the sensing-and-learning chain, and what additional information is required before that failure can be repaired.

The RQ answers above delimit the central finding: familiar-condition learnability does not translate into useful Strict-DG transfer to P4 within the principal campaign, and the tested remedies do not supply a reliable general rescue.

The measurement campaigns analysed in this dissertation pre-date the present work. The original contribution of the dissertation lies in the governed evaluation design, source-only selection analysis, diagnostic and intervention experiments, reproducibility controls, and integrated interpretation applied to those measurements.

These conclusions are deliberately bounded. They do not establish that domain generalisation is universally impossible in chipless RFID, that one electromagnetic mechanism uniquely causes the P4 failure, that every tested method family fails in other designs or datasets, or that the exploratory OpenEMS branch provides a fabrication-ready redesign. The boundaries specify the domain of validity of the conclusions rather than weakening the evidence within that domain.

6.3.3 Priorities for Future Work

The next work should be prioritised according to which unresolved question it can remove, rather than by continuing to search the existing P4 corpus for another loss function.

First, perform a prospective held-geometry replication. The highest-value experiment is a new acquisition campaign in which the target geometry is fixed before model development and target scoring. Ideally, more than one held geometry should be used and distance and angle should be sampled at more than two values. The same condition-block discipline should be retained, and the independently acquired target should not be reused for source-side method selection. This would test whether the present failure and angle association reproduce beyond one held joint geometry.

Second, perform a causal representation intervention. The present position probe shows strong source-position decodability, but DANN did not reliably suppress it. A stronger test would first demonstrate, using source data only, that an intervention materially reduces position decodability while retaining source TagID discrimination. Only after that mechanism check is satisfied should the frozen representation be evaluated on a prospectively held geometry. An improvement under that matched intervention would provide stronger evidence about whether geometry-entangled representation learning contributes causally to the transfer failure.

Third, acquire richer physical measurements and validate physical redesign directly. A future campaign should preserve the authoritative frequency axis and, where practical, retain complex response and richer polarimetric information. Those measurements would allow the signal-level hypotheses raised by the present magnitude-only diagnostics to be tested more directly. Any OpenEMS redesign should first satisfy its simulation confirmation criteria, then be independently checked against the complete measurement chain, and only

then proceed to fabrication and re-measurement under the same held-geometry evaluation discipline.

A common-protocol multi-geometry benchmark would be a valuable longer-term extension if sufficient independent campaigns become available. Its value would be methodological: shared condition-block definitions, declared target-access rules and genuinely held geometries would allow future algorithmic improvements to be distinguished from changes in the evaluation regime.

Familiar-condition success is insufficient evidence for a chipless-RFID reader expected to operate across installations. In this campaign, 45° acquisition—shared by P2 and the held P4 geometry—is associated with poorer TagID discrimination than 0° acquisition at the measured-signal level, while at P4 the source-trained representation preserves still less linearly accessible discrimination than RAW. The tested source-only objectives and limited target-access interventions do not provide a reliable general rescue. Robust deployment therefore requires acquisition design, representation design and evaluation protocol to be developed together, rather than expecting the classifier alone to absorb the full measurement shift.

References

[1] S. Preradovic and N. C. Karmakar, "Chipless RFID: Bar code of the future," IEEE Microw. Mag., vol. 11, no. 7, pp. 87–97, 2010, doi: 10.1109/MMM.2010.938571.

[2] C. Herrojo, F. Paredes, J. Mata-Contreras, and F. Martín, "Chipless-RFID: A review and recent developments," Sensors, vol. 19, no. 15, Art. no. 3385, 2019, doi: 10.3390/s19153385.

[3] N. Rather, R. B. V. B. Simorangkir, J. L. Buckley, B. O'Flynn, and S. Tedesco, "Deep-learning-assisted robust detection techniques for a chipless RFID sensor tag," IEEE Trans. Instrum. Meas., vol. 73, pp. 1–10, Art. no. 2502710, 2024, doi: 10.1109/TIM.2023.3334378.

[4] N. Rather, R. B. V. B. Simorangkir, D. R. Gawade, J. L. Buckley, B. O'Flynn, and S. Tedesco, "Hybrid DCNN-enabled depolarizing chipless RFID: Improving tag detection across varying lossy surfaces and shapes," IEEE J. Radio Freq. Identif., vol. 9, pp. 768–778, 2025, doi: 10.1109/JRFID.2025.3608617.

[5] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 4, pp. 4396–4415, 2023, doi: 10.1109/TPAMI.2022.3195549.

[6] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[7] S. Preradovic, I. Balbin, N. C. Karmakar, and G. F. Swiegers, "Multiresonator-based chipless RFID system for low-cost item tracking," IEEE Trans. Microw. Theory Techn., vol. 57, no. 5, pp. 1411–1419, 2009, doi: 10.1109/TMTT.2009.2017323.

[8] A. Vena, E. Perret, and S. Tedjini, "Chipless RFID tag using hybrid coding technique," IEEE Trans. Microw. Theory Techn., vol. 59, no. 12, pp. 3356–3364, 2011, doi: 10.1109/TMTT.2011.2171001.

[9] F. Costa, S. Genovesi, and A. Monorchio, "A chipless RFID based on multiresonant high-impedance surfaces," IEEE Trans. Microw. Theory Techn., vol. 61, no. 1, pp. 146–153, 2013, doi: 10.1109/TMTT.2012.2227777.

[10] D. P. Mishra and S. K. Behera, "Resonator based chipless RFID: A frequency domain comprehensive review," IEEE Trans. Instrum. Meas., vol. 72, pp. 1–16, Art. no. 5500716, 2023, doi: 10.1109/TIM.2022.3225027.

[11] A. Subrahmannian and S. K. Behera, "Chipless RFID: A unique technology for mankind," IEEE J. Radio Freq. Identif., vol. 6, pp. 151–163, 2022, doi: 10.1109/JRFID.2022.3146902.

[12] A. Vena, E. Perret, and S. Tedjini, "A depolarizing chipless RFID tag for robust detection and its FCC compliant UWB reading system," IEEE Trans. Microw. Theory Techn., vol. 61, no. 8, pp. 2982–2994, 2013, doi: 10.1109/TMTT.2013.2267748.

[13] A. Ramos, Z. Ali, A. Vena, M. Garbati, and E. Perret, "Single-layer, flexible, and depolarizing chipless RFID tags," IEEE Access, vol. 8, pp. 72929–72941, 2020, doi: 10.1109/ACCESS.2020.2988116.

[14] F. Costa, S. Genovesi, and A. Monorchio, "Normalization-free chipless RFIDs by using dual-polarized interrogation," IEEE Trans. Microw. Theory Techn., vol. 64, no. 1, pp. 310–318, 2016, doi: 10.1109/TMTT.2015.2504529.

[15] A. Lazaro, A. Ramos, D. Girbau, and R. Villarino, "Chipless UWB RFID tag detection using continuous wavelet transform," IEEE Antennas Wireless Propag. Lett., vol. 10, pp. 520–523, 2011, doi: 10.1109/LAWP.2011.2157299.

[16] J. J. Fodop Sokoudjou, F. Villa-González, P. García-Cardarelli, J. Díaz, D. Valderas, and I. Ochoa, "Chipless RFID tag implementation and machine-learning workflow for robust identification," IEEE Trans. Microw. Theory Techn., vol. 71, no. 12, pp. 5147–5159, 2023, doi: 10.1109/TMTT.2023.3276011.

[17] F. Villa-González, J. J. Fodop Sokoudjou, O. Pedrosa, D. Valderas, and I. Ochoa, "Analysis of machine learning algorithms for USRP-based smart chipless RFID readers," in Proc. 17th Eur. Conf. Antennas Propag. (EuCAP), Florence, Italy, 2023, pp. 1–5, doi: 10.23919/EuCAP57121.2023.10133472.

[18] J. J. Fodop Sokoudjou, P. García-Cardarelli, A. Rezola, D. Valderas, J. Díaz, and I. Ochoa, "Chipless RFID tag detection based on continuous wavelet transform and convolutional neural networks," IEEE Trans. Microw. Theory Techn., vol. 73, no. 9, pp. 6260–6274, 2025, doi: 10.1109/TMTT.2025.3559537.

[19] N. Rather, R. B. V. B. Simorangkir, J. L. Buckley, B. O'Flynn, and S. Tedesco, "1D-CNN enabled depolarizing chipless RFID," in Proc. IEEE Int. Conf. RFID Technol. Appl. (RFID-TA), Valence, France, 2025, pp. 1–4, doi: 10.1109/RFID-TA63091.2025.11265818.

[20] J. J. Fodop Sokoudjou, P. García-Cardarelli, A. Rezola, E. Perret, and I. Ochoa, "AI for chipless RFID tag detection: Generalization and deployability on embedded digital signal processors," in Proc. IEEE Int. Conf. RFID Technol. Appl. (RFID-TA), Valence, France, 2025, pp. 1–4, doi: 10.1109/RFID-TA63091.2025.11265846.

[21] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," IEEE Trans. Knowl. Data Eng., vol. 35, no. 8, pp. 8052–8072, 2023, doi: 10.1109/TKDE.2022.3178128.

[22] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in Computer Vision – ECCV 2016 Workshops, LNCS vol. 9915, 2016, pp. 443–450, doi: 10.1007/978-3-319-49409-8_35.

[23] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," J. Mach. Learn. Res., vol. 13, no. 25, pp. 723–773, 2012.

[24] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, no. 59, pp. 1–35, 2016.

[25] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," arXiv:1907.02893, 2019. [Online]. Available: https://arxiv.org/abs/1907.02893.

[26] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.

[27] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.

[28] Y. Wu and K. He, "Group normalization," in Computer Vision – ECCV 2018, LNCS vol. 11217, 2018, pp. 3–19, doi: 10.1007/978-3-030-01261-8_1.

[29] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-based image classification: Generalizing to new classes at near-zero cost," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 11, pp. 2624–2637, 2013, doi: 10.1109/TPAMI.2013.83.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.

[31] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, no. 86, pp. 2579–2605, 2008.

[32] J. J. Fodop Sokoudjou, "Chipless RFID measurement for ML predictions," IEEE DataPort, 2022, doi: 10.21227/681r-eb17.

[33] K. S. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," IEEE Trans. Antennas Propag., vol. 14, no. 3, pp. 302–307, 1966, doi: 10.1109/TAP.1966.1138693.

[34] T. Liebig, A. Rennings, S. Held, and D. Erni, "openEMS—A free and open source equivalent-circuit (EC) FDTD simulation platform supporting cylindrical coordinates suitable for the analysis of traveling wave MRI applications," Int. J. Numer. Model., Electron. Netw. Devices Fields, vol. 26, no. 6, pp. 680–696, 2013, doi: 10.1002/jnm.1875.

Appendix A — Study Register and Evidence Map

Table 28 identifies each study, the information used during development and the conclusion it supports. Appendix D summarises verification and access controls; detailed evidence identities are retained in the final frozen release manifest.

Table 28. Study register and evidence map. Appendix D summarises verification; detailed file identities are retained in the final frozen release manifest.

Study	Loc.	Access class	Final outcome
Historical grouped/random work	§4.1	Familiar/mixed-condition historical	Learnability context only; not Strict-DG.
Fixed-position grouped benchmark	§4.3	Target-domain diagnostic	P4 not uniquely hardest; P2 comparably weak under this protocol.
Synchronised angle-distance factorial	§4.3	Physical/signal diagnostic	Angle association supported; dominance is not established. Distance main effect unconfirmed; the negative 0° simple effect is supported for Macro-F1, but not seed-inclusive Accuracy. Interaction floor-limited.
Primary first-difference source-only evaluation	§4.2	Strict source-only DG	Accuracy 0.156635; Macro-F1 0.112231 over five seeds.
Source-selection regret	§4.2	Source-only diagnostic	C1 oracle 2/3; mean regret 0.013439; descriptive source-score reuse support, not nested validation.
Representation-versus-signal	§4.4	Target-domain diagnostic	RAW exceeds C1 at P4 by 0.0951 [0.0173, 0.1747].
Position diagnostic	§4.4	Source-only/target diagnostic	Position strongly decodable; causal role unresolved.
Physical/spectral analysis	§4.4	Physical/signal diagnostic	Localised associative evidence; 281-point crop; global missingness gate failed, but missingness was lowest at P4 and does not explain transfer failure.
Historical 1/3/5-shot	§5.2	Few-shot target adaptation	3-shot descriptive gain on one reused P4 corpus.
Factor-aware Shot-P4	§5.2	Few-shot target adaptation	No reliable joint ER/surface support advantage.
Large P4 calibration	§5.2	Target-assisted	Nonmonotonic; peak 0.135560 at 20 labels, 0.102347 at 500.
Trainable linear readout	§5.2	Target-assisted diagnostic	0.142740 at 500; modest uncertain point recovery, no rescue.
Encoder fine-tuning	§5.2	Target-assisted diagnostic	Partial/full below frozen linear; target-label ladder closed.
Retrospective P4 reference	§5.3	Retrospective target-informed	0.577143/0.583569, n=1, different historical representation path; never Strict-DG.
Matched A0/A1	§5.3	Target-assisted validation selection	A1 selects frozen source-fitted readouts on P4-Val; P4-Test labels are protected until selection freeze. Small uncertain effect within represented conditions.
Matched A2	§5.3	Within-condition interpolation	0.983494 dependence-inflated; leave-condition controls collapse.

Study	Loc.	Access class	Final outcome
Data1 portability	App. E	External portability	Strong within-Data1 classification only.
External Data1 SSL	App. E	External SSL / source-only DG	No reliable or practical cross-polar depolarizing-tag dataset P4 benefit.
IRM	§5.1	Strict source-only DG	Valid implementation; no reliable matched benefit.
DANN	§5.1	Strict source-only DG	No reliable position suppression or matched P4 benefit.
GroupDRO	§5.1	Strict source-only DG	Metrics recovered from unchanged stored predictions; no reliable benefit.
Condition-matched cross-position Mixup	§5.1	Source-only feasibility diagnostic	50% eligible-pair coverage versus the >99% feasibility threshold; zero runs; no rerun.
Unified expanded DG benchmark	§5.1	Matched-effect synthesis	Matched comparison completed; interpretation subject to stated limitations.
OpenEMS redesign	App. F	Simulation / design exploration	NOT_CONFIRMED; 0.681% change; fabrication not recommended.
OpenEMS R2	App. F	Simulation / design exploration	No simulations were executed; no additional scientific result.

The primary source-only predictions, method-specific studies and historical baseline are retained in separate project records. Their detailed identities and hashes belong to the final frozen release manifest; Appendix D summarises verification without implying a shared development history.

Appendix B — Protocol and Hyper-parameter Detail

B.1 Strict source-only recipe

Element	Setting
Input representation	280-point first difference of the 281-point spectrum
Normalisation	source-only feature-wise population standardisation; never refitted on target
Encoder	1-D CNN: channels 64/128/256, kernels 7/5/3; ReLU, two max-pool layers, adaptive average pooling, 256-D embedding, seven-class linear head and GroupNorm (8 groups) [28]
Optimiser	Learning rate 0.001; weight decay 0.0001; AdamW [30]
Batch size	256
Loss	unweighted cross-entropy
Class weighting	none
Label smoothing	0
Learning-rate scheduler	none
Seeds	42, 43, 44, 45, 46 — all five reported, none discarded
Final epochs by seed	13, 13, 9, 10, 9; source selection permits up to 50 epochs with patience 8, followed by the frozen final epoch rule
Selection criterion	lexicographic; primary criterion: worst held-source Macro-F1 across P1–P3
Source-selection units	4 candidates × 3 folds × 5 seeds = 60
Target access	inference only, after recipe freeze and checkpoint verification
Held-source regret	For each held-source event, the selector reaggregates stored score rows from the other two permitted source positions; all choices are frozen before held-event scoring. The 60 stored units are reused without retraining. This is a source-score reuse diagnostic, not a fresh nested model-development experiment.

B.2 Fixed-position and factorial protocol

Element	Setting
Design	4 positions × 3 outer folds × 5 seeds = 60 primary runs (Section 4.3)
Fold construction	Latin-square outer folds: 21 held blocks (1,050 rows), 28 training blocks and 14 validation blocks per position; refit uses the 42 non-held blocks.
Primary endpoint	Table 11 reports row metrics, with held rows pooled across the 63 blocks per seed for Macro-F1. The synchronised factorial study uses condition-aggregated predictions as specified below.
Uncertainty	10,000-replicate paired, tag-stratified block bootstrap; seeds fixed
Sensitivity	the same bootstrap with paired seed resampling
Factorial amendment	60 synchronised reruns with position-independent validation assignment, frozen before training; inherited runs used only for reconstruction
Learning validity	56-row true-label subset fitted at every position without changing primary hyper-parameters

B.3 Representation diagnostic

Element	Setting
Representations	RAW_SIGNAL_281, FIRST_DIFFERENCE_280, C1_ENCODER_EMBEDDING (256-d, network.12), PEAK_DESCRIPTOR
Probe	single linear layer with softmax; Adam, learning rate 0.01, weight decay 1×10^{-4}; Xavier-uniform weights, zero bias (Section 4.4)
Training	full batch up to 2,048; at most 200 epochs; validation-macro-F1 selection; patience 20
Seeds	42–46, with encoder checkpoint seeds aligned to probe seeds
Endpoint	macro-F1 over 63 out-of-fold condition-block majority predictions per position and seed
Peak windows	Supporting only: extraction-validity gate failed; definitions and missingness are in Appendix G.2 (see §4.4.5).

B.4 Few-shot protocol

Element	Setting
Representation	frozen 256-d penultimate output of the source-only C1 encoder; encoder, head and preprocessing all frozen
Episodes	18 pre-specified episodes from two cyclic families over the nine (surface, permittivity) combinations
Support / query	5 combinations form the support pool; 4 form a 1,400-sample query set identical across shots within an episode; all support-pool blocks, including unselected blocks, are excluded from query evaluation (Section 5.2.1).
Budgets	0, 7, 21, 35 labels for 0-, 1-, 3-, 5-shot; class-balanced and nested
Support ranking	SHA-256 over a frozen salt and identity fields; signal magnitude and model output excluded by construction
Adapters	target-only squared-Euclidean prototypes; source-anchored linear head refitted by L-BFGS under a proximal penalty
Penalty selection	chosen from {0.01, 0.1, 1, 10, 100} on source pseudo-target evidence before any target execution; selected 100, 100 and 0.1
Aggregation	mean over five seeds within an episode, then equal-weight mean over 18 episodes
Uncertainty	paired-episode percentile bootstrap, 10,000 resamples

B.5 Extended objectives

Method / diagnostic	Development or support	Selected setting	Final interpretation
IRM	75 development; 10 final	lambda 1	-0.019330 [-0.066939, +0.025387] vs matched ERM
DANN	source-only development; 15 frozen predictions	lambda 0.1	+0.016037 [-0.035578, +0.068624]; no reliable suppression
GroupDRO	75 development; 10 final	eta 0.05	+0.001405 [-0.042978, +0.039489]; persistence recovery disclosed

Method / diagnostic	Development or support	Selected setting	Final interpretation
Condition-matched cross-position Mixup	0 development; 0 final; 0 P4	not reached	50% eligible-pair coverage; final feasibility failure
Labelled-target ladder	budgets through 500	prototype, linear, partial/full	nonmonotonic calibration; no readout or encoder rescue
Prototype label ladder	Nested prefixes; 0/7/10/20/21/35/50/100/200/500 labels per fold	Fixed cosine prototypes; frozen source preprocessing and encoder; no target validation search	3 fixed query folds × 5 source seeds × 20 support selections; row predictions pooled out of fold before seed/support averaging
Ladder condition coverage	21 query blocks and 42 available support blocks per fold	Budgets 7–35 cover that many blocks; budgets ≥50 cover all 42	500 labels include 458 sweeps beyond the first per support block; repeated spectra remain possible within support
Ladder query protection	Support/query condition, row and exact-signal disjointness within each fold	Query labels: final scoring only; all 3,150 rows queried out of fold	A block may support another fold; folds do not constitute independent target corpora
Ladder uncertainty	10,000 paired TagID-stratified condition-block replicates	Crossed source-seed/support-selection resampling	Label count and block coverage co-vary below saturation; another calibration design is not excluded

Appendix C — Metric Definitions and Aggregation Rules

Different studies use different statistical units. Table 29 records the applicable rule; no single row-level or block-only interval is treated as universally governing seeds, folds, episodes, selection events and target partitions.

Table 29. Aggregation rules by study family.

Quantity	Definition and aggregation
Row Accuracy	Fraction of measurement rows classified correctly. A row is a scoring item, not an independent inferential unit when sweeps share a condition; uncertainty then resamples physical condition blocks.
Majority-vote condition-block Accuracy	Fraction of condition blocks correctly classified after majority voting over row predictions within each block. Exact ties resolve to the lowest class index. This is distinct from averaging row correctness within a block; study-specific aggregation must be stated.
Macro-F1 (per run)	unweighted mean of the seven per-class F1 values. A class with no predictions contributes zero rather than an undefined value.
Pooled metrics across condition blocks	The scoring item and pooling unit are separate. Table 11 pools held measurement rows from all 63 blocks within each seed before computing Macro-F1 and averaging seeds. Metrics computed after one prediction per block instead use condition-aggregated predictions. Nonlinear Macro-F1 differs from the mean of fold-level Macro-F1.
Population SD	ddof = 0 unless stated otherwise. A zero SD from a single run is arithmetic, not evidence of stability.
Bootstrap interval	10,000 deterministic replicates, paired and tag-stratified over condition blocks, 95% percentile, fixed random seed.
Seed-sensitivity interval	as above, additionally resampling training seeds. Reported wherever it changes the conclusion; the more conservative interval governs.
Balanced uniform-random reference	For the balanced seven-class task, uniform random prediction has expected Accuracy and Macro-F1 of approximately 1/7. A majority-class predictor also has Accuracy 1/7, but its Macro-F1 is lower because the remaining six classes have zero recall.
Block resolution	Row Accuracy changes by 50 / 3150 = 0.0159 per fully reclassified condition block; this is a useful physical reference scale, not a formal significance threshold.
Within-condition row Accuracy	For Figure 17, each physical condition contributes 50 sweeps × 5 seeds = 250 predictions. Its cell value is correct predictions / 250 (the printed annotation is the numerator). This is a descriptive within-condition row statistic, not majority-vote condition-block Accuracy; pooled row Accuracy remains 0.1566.
Mean-logit condition-block metrics	The frozen fixed-position implementation and synchronised factorial study first average the logits over the 50 sweeps within a condition and choose the highest mean logit, then score one prediction per block. This differs from majority voting over row predictions. Table 12 uses this condition-level aggregation; Table 11 uses row metrics.

Appendix D — Reproducibility Statement

Table 30. Evidence level by study. 'Retrained' means model fitting was executed again; 'replayed' means stored predictions were re-scored by independent code without refitting.

Study	Evidence level	Notes
Strict source-only — final models	retrained	All five models retrained in the locked environment; all five target logit and prediction arrays reproduced exactly. Independently re-derived from raw files by separate code, matching bitwise.
Strict source-only — selection	replayed	60 units hash-verified where evidence permits, independently re-scored and re-ranked. Full 60-model retraining is not claimed.
Fixed-position benchmark	executed	60 primary runs executed under ten pre-training gates.
Angle–distance factorial	executed	Inherited runs reproduced released metrics exactly; 60 synchronised runs executed for inference.
Representation diagnostic	executed	Evaluation predictions hashed before the corresponding evaluation labels were opened; authorised fitting/selection labels are governed separately. Five seeded controls passed for every active representation/position pair.
Source-selection regret	replayed	60 units replayed from hash-bound prediction bundles; no retraining; target seal intact.
Few-shot	executed	Support/query disjointness and query-label sealing independently verified; 54 isolation checks passed.
Retrospective analysis	partially replayable	Metrics replay exactly from stored predictions within the archive; regenerating predictions from the representation requires an externally retained hash-verified carrier.
Matched comparison	executed	A0 uses source fitting/selection; A1 uses P4-Val selection among frozen readouts; A2 fits on P4-Adapt and selects on P4-Val. The selection registry is hashed before P4-Test scoring. All arms share physical conditions across roles; 27 independent audit tests passed.
Extended objectives	executed / feasibility-only	IRM, DANN and GroupDRO executed under their branch protocols. Mixup stopped before training because the pre-specified feasibility criterion was not met; GroupDRO metrics were recovered from unchanged frozen predictions.
Redesign simulation	executed with limitations	Pilot-to-confirmation agreement poor relative to the figure of merit.

Table 30 shows that frozen results retain their recorded environments, seeds, predictions and manifests. The final release manifest resolves all canonical evidence paths and binds the historical baseline and disjoint patch modules separately. This reconstruction changes no scientific artifact and creates no common Git ancestry.

Scientific freeze. The scientific content of this dissertation is frozen to the repository evidence available on 10 August 2026. No model was trained, retrained or extended for this

revision; all reported results are transcriptions of frozen artifacts under their governing access class. The canonical historical baseline and the later patch modules have separate repository roots, and the release manifest binds their commits and evidence hashes separately rather than asserting a common ancestry. The short historical recipe digest did not cover the shuffle-stage offset, convolution padding, activation, pooling, early-stopping settings or CORAL weight. Those details are retained in the full release specification, code, configuration, checkpoints and prediction digests.

Historical baseline reproduction. The historical pipeline was rerun in the current environment. Fifty-three of 54 per-run accuracy comparisons and all 18 split means were within the pre-specified tolerance of 0.05. The single exception was the co-polar dataset's leave-surface BEND2 run at seed 44: 0.4984 versus 0.5495, a difference of 0.0510. Historical predictions and checkpoints were unavailable, so this comparison is based on retrained metrics rather than binary identity.

The honest summary is therefore that the historical baselines are substantially but not completely reproduced, that the failure is confined to one run of the secondary dataset, and that the headline grouped-random and leave-position figures used in Section 4.1 are within tolerance. They are used only as context and no conclusion in this dissertation depends on them.

D.1 Research Infrastructure and Scientific Governance

This appendix records how the experimental work was organised and how the information boundaries in Chapter 1 were enforced. The main text retains only the methodological rule and the evidence needed to interpret the results.

D.1.1 Why the infrastructure was rebuilt

The project accumulated several independent working trees, each with its own conventions, status documents and archived results. Numbers produced under different information boundaries came to sit beside one another in the same summary tables. In at least one case a result that had been selected using target-domain outcomes was described in project documentation as a source-only benchmark. Neither was a computational error; both arose from inadequate control of information access and result provenance, which can turn a defensible result into an indefensible claim.

The final repository was therefore organised with the explicit aim of making the information boundary a property of the code rather than of the narrative. Each scientific study has its own entry point, its own configuration, its own data-access contract and its own claim-

boundary record, and a repository-level status file states what each study may and may not be used to assert.

D.1.2 Mechanisms that enforce the boundary

- **Authorisation-gated target loading.** In the strict source-only workflow the target position cannot be loaded until the recipe, the preprocessing state and all five model checkpoints have been re-derived and verified from disk, and target labels are released only after predictions have been serialised.
- **Hashed preregistration.** Protocols, hyper-parameter grids, selection rules and uncertainty procedures are written and hashed before execution. Where a protocol had to be amended mid-study, the amendment is recorded with its own timestamp and the reason.
- Prediction freezing and label access. Label access is defined separately for fitting, selection and evaluation. Strict-DG excludes P4 labels from fitting and selection. Target-assisted and diagnostic studies may use only the support or validation labels explicitly authorised by their protocols. Prediction-freezing statements therefore refer to the evaluation or query labels protected by the corresponding study, rather than to every target label. In matched A0/A1, P4-Val labels may select A1's frozen source-fitted readout; the selection registry is frozen and hashed before P4-Test labels are scored. Authorised support/validation access does not authorise adapting to protected evaluation outcomes.
- **Read-only isolation.** Each later study runs in its own isolated copy of the repository. On completion the copy verifies that the authoritative trees and all earlier studies are byte-identical to their baselines, so no study can silently modify another's evidence.
- **Explicit claim registers.** Every study records the claims it supports, the claims it forbids, and the evidence path for each. The forbidden-claim lists are the operative part: they name the specific over-reading that the evidence does not support.

D.1.3 Replay versus retraining

One distinction recurs in the reproducibility statements and is worth defining once. **Retraining** means that model fitting was executed again from the recipe and the resulting parameters compared. **Replay** means that stored predictions were re-scored by independent code without refitting the model. Replay verifies that reported metrics follow from stored predictions; it does not verify that those predictions follow from the recipe.

Both appear in this work and are labelled wherever they matter. In the strict source-only study, for example, the sixty source-selection units were replayed and independently re-scored but not retrained, whereas the five final models were retrained in a locked

environment and reproduced their stored predictions exactly. Section 4.2 states this explicitly rather than describing the whole study as reproduced.

D.1.4 Independent verification

The principal result of Section 4.2 was checked by an independent procedure that did not use the project's inference or metric code. That procedure parsed the target position directly from the three raw measurement files, re-declared the network from the recipe text, and recomputed the metrics. All five seeds matched bitwise on both logits and predictions, and every metric matched exactly; retraining from a re-declaration of the training procedure produced bitwise-identical model states. The comparison is not circular, because the same per-seed prediction and logit digests appear in a pre-existing table produced three days earlier by different code.

Several later studies were also reviewed independently with access to the isolated repository but not to the analysis narrative. Where such an audit disagreed with the study's own account, both are reported; Section 5.3.3 records the historical-carrier selector discrepancy.

D.1.5 Status vocabulary used in this dissertation

The repository records study outcomes with a controlled vocabulary. That vocabulary is internal and is deliberately not used in the narrative of this dissertation, which uses ordinary scientific language instead. Table 31 maps this vocabulary for comparison with the repository.

Table 31. Correspondence between internal status vocabulary and the language used in this dissertation.

Internal status	Meaning as written here
NEGATIVE_RESULT	The study completed its protocol but did not establish the hypothesised benefit, or recorded an effect in the unfavourable direction; this does not by itself establish a zero effect.
NOT_CONFIRMED	the study completed but its recorded confirmation criteria were not met
PASS_WITH_LIMITATIONS	the study produced usable evidence subject to stated limitations
BLOCKED	the study could not be executed; no scientific result is reported
HISTORICAL_REFERENCE_ONLY	retained for context; not evidence for a current claim
EXECUTION_COMPLETE_WITH_LIMITATIONS	the protocol ran to completion but some component was replayed rather than re-executed, or a documented deviation occurred

The most important translation is the first. A negative result is a completed scientific outcome, not a failed experiment. Several of the principal findings of this dissertation are negative results in exactly this sense, and they are reported as findings.

Appendix E — External Dataset and Self-Supervised Checks

Two separate studies use an independently acquired public chipless-RFID corpus [32]. They ask different questions, reach different conclusions and should not be conflated.

E.1 Why cross-dataset work is constrained here

The obvious use of a second dataset would be to test whether a model trained on the principal dataset works on it. That test is not available, and the reason is structural rather than practical. Two tag sets designed by different groups do not encode the same identities: the external corpus has four classes with their own geometry and their own encoding band, and there is no mapping under which its class 2 corresponds to any class of the seven-tag depolarizing set. There is no shared label space, so there is nothing for a seven-class classifier to predict.

What a second dataset *can* support is a test of whether the processing, splitting, evaluation and uncertainty machinery transfers — and a test of whether unlabelled external measurements can help the principal problem through representation learning. Those are the two studies reported below.

E.2 Study 1: pipeline portability within the external corpus

The first study applies the same custody, grouping, training and evaluation code to the external four-class corpus, with no reference to the principal dataset. Table 32 and Figure 14 summarise the results.

Table 32. Within-corpus results on the external four-class dataset. CNN figures are equal-weight means over five seeds.

Method	Runs	Accuracy	Macro-F1
Majority-class baseline	1	0.297297	0.114583
PCA + logistic regression	1	0.996216	0.994823
Raw-signal CNN	5	0.748865	0.678840
First-difference CNN	5	0.693081	0.607360

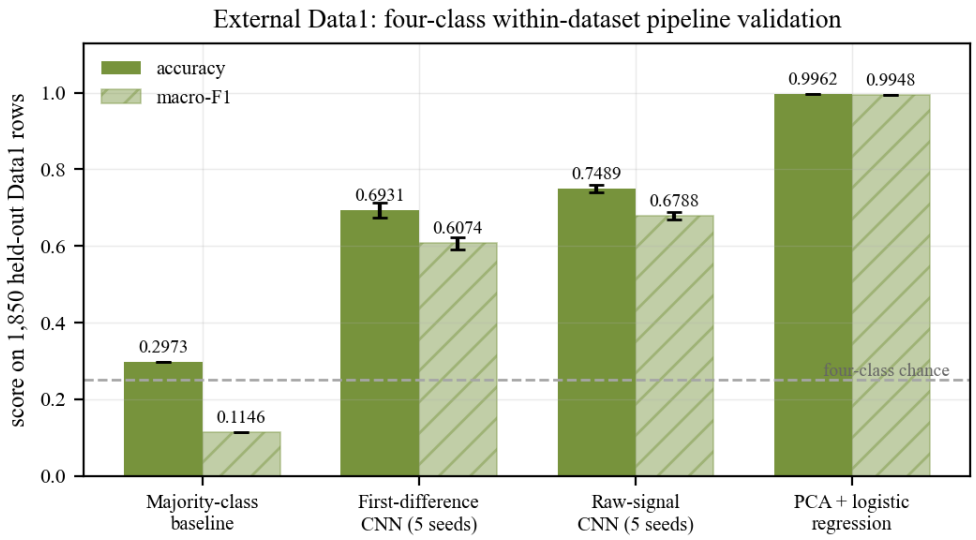

Figure 14. Within-corpus performance on the external four-class Data1 corpus. This supports scoped pipeline portability only; it does not validate the seven-class model or cross-position transfer. The dashed line is the uniform-random Accuracy reference = 0.25, not a calibrated Macro-F1 chance reference. CNN error bars are descriptive (see note).

Note. CNN bars show five-seed means. The retained CNN error bars have no verified SD/CI definition and are not used for inference.

Within this four-class corpus, a linear model on principal components reaches 0.996, far above either convolutional network. That says something about the external corpus — its classes are close to linearly separable in the raw feature space — and nothing about the principal problem, where no method approaches such values across positions.

E.3 Study 2: can unlabelled external data help?

The second study asks a question that could in principle help the principal problem: does self-supervised pre-training on additional unlabelled measurements — from the principal dataset's source positions, from the external corpus, or from both — improve strict source-only performance at the held-out position? The intuition is reasonable. If the encoder collapses at the target because it has seen too few acquisition conditions, then more unlabelled spectra might teach it a more general notion of what a chipless-RFID spectrum looks like.

Table 33 summarises the five treatments fixed before target evaluation; P4 was scored once under the same rule.

Table 33. Self-supervised treatments under the source-only P4 protocol. T0 is the method-specific supervised control and is not numerically identical to the primary result in Table 10.

Treatment	Accuracy	Macro-F1
T0 matched supervised-only control	0.158984	0.123939
T1 principal-dataset source-only SSL	0.135238	0.103003

Treatment	Accuracy	Macro-F1
T2 external-corpus-only SSL	0.144317	0.104029
T3 joint balanced SSL	0.143683	0.106758
T4 joint physics-aware SSL	0.133270	0.109105

The planned external-data contrast compares learned-head Macro-F1 for T3 joint balanced SSL minus T1 principal-dataset source-only SSL, averaged equally over the five paired source seeds. It tests the added external-data component while retaining SSL, rather than comparing each SSL arm with supervised-only T0. The recorded mean contrast is +0.003755, positive in only 2 of 5 seeds and below the declared practical threshold of 0.02.

No self-supervised treatment beat the matched supervised-only control: T0 reached Macro-F1 0.123939, above the best self-supervised head at 0.109105. Under the tested source-only SSL protocol, unlabelled Data1 did not establish a reliable or practically meaningful P4 benefit. This is not target adaptation and does not show that Data1 improved P4 transfer.

Appendix F — Electromagnetic Design Exploration

The diagnostic results motivate asking whether tag geometry could improve spectral separation, but they do not establish that a redesign would improve P4 classification. This appendix is exploratory physical design work and is kept separate from the reported recognition evidence.

F.1 The chain of reasoning

The argument that motivates a redesign runs as follows.

- Transfer to the unobserved position fails, and it fails by class collapse rather than by uniform degradation (§4.2.4).
- The degradation is present in the measured spectrum before encoding, and is associated with reader angle (§4.3.3, §4.4.3).
- The tag code is carried by a small number of resonances whose depth and visibility change with geometry (§4.4.5).
- Under the tested four-peak descriptor, only 13 of 252 tag-pair × peak × position combinations were reliably separated (Appendix G.2). This indicates limited separability under that descriptor, but it is not a direct measurement of physical codebook margin.
- Therefore a geometry that increases the worst-case spectral separation between tags should be more robust to the loss of shallow features.

The final step is a design hypothesis evaluated only in simulation. It is not a hardware, RF-measurement or classification intervention.

F.2 Simulation setup and its assumption boundary

Candidate geometries were evaluated by finite-difference time-domain simulation [33], using the openEMS solver [34]. The figure of merit is the **minimum separation across all tag pairs** — the worst case, not the average — because the failure mode being addressed is the loss of the least distinguishable pairs. A pilot stage screened candidates and a confirmation stage re-simulated survivors under tighter conditions, with agreement between the two stages required before a candidate could be recommended.

The assumption boundary must be stated because it bounds every number in this section. The simulation models the tag and its immediate substrate. It does not model the measurement chain that produced the dataset: the reader antennas, their patterns, the cross-polar receive path, the 50 mm and 150 mm geometries or the mounting surfaces. A separation improvement in simulation is therefore a statement about the tag's intrinsic

scattering, and the inference from that to classification performance under position shift is an assumption, not a measurement.

F.3 Result

Table 34. Exploratory OpenEMS design evidence. Confirmation criteria were not met, and no hardware or classification validation is claimed.

Quantity	Value	Interpretation
Baseline minimum separation	3.192895 dB	existing simulated geometry
Candidate minimum separation	3.214639 dB	best recorded candidate
Relative improvement	0.681%	below recorded 5% practical gate
Pilot-confirmation RMS	7.028 dB	confirmation mismatch
Maximum peak drift	550 MHz	confirmation mismatch
Bottleneck pair	110-111 to 100-110	descriptive relocation
Outcome	Confirmation criteria not met	no hardware or classification validation
Fabrication recommendation	Not recommended	publication extension only after new evidence

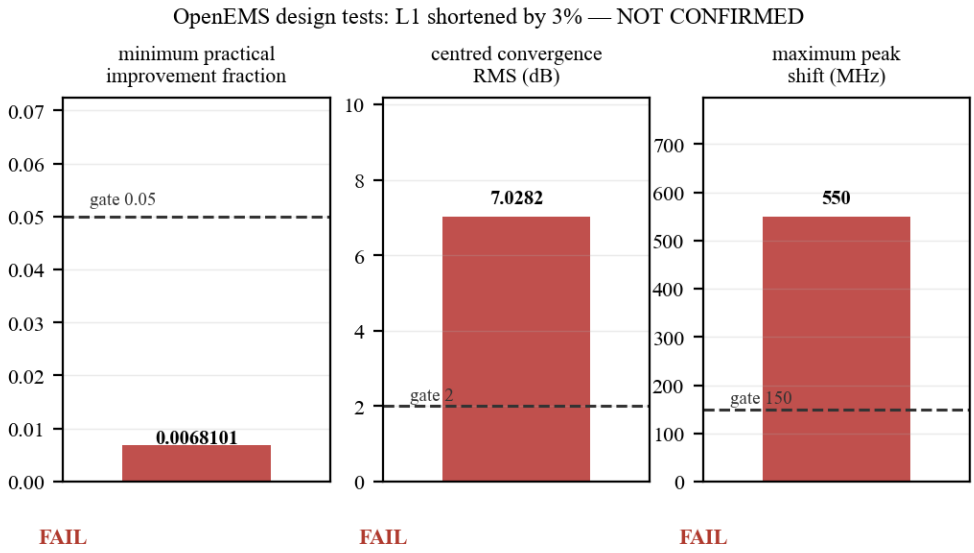

Figure 15. Recorded OpenEMS design tests and outcomes. The study did not meet its confirmation criteria and no hardware validation is claimed.

The study is **not confirmed** and **fabrication is not recommended**. Three aspects of the result in Table 34 and Figure 15 are worth separating.

The candidate raises the recorded minimum separation from 3.192895 dB to 3.214639 dB, a relative improvement of 0.681%, below the recorded 5% practical gate. No classification consequence can be inferred from this point change.

Pilot-to-confirmation RMS mismatch is 7.028 dB and maximum peak drift is 550 MHz. These values fail the recorded confirmation gates. The evidence does not identify meshing

as the definitive cause, and the exact candidate universe, ranking configuration and timing of threshold specification remain unresolved.

The bottleneck pair changes from 110-111 to 100-110. This is a descriptive simulation outcome, not proof that the codebook is at a physical limit.

F.4 The robustness follow-up

No OpenEMS R2 simulations were executed, so this follow-up contributes no additional scientific result. A measured or fabricated follow-up would test the design hypothesis.

F.5 What would be required for a manufacturing recommendation

- A simulation whose pilot-to-confirmation agreement is small relative to the figure of merit — the current 7.03 dB against 3.2 dB is the first thing to fix.
- A model of the measurement chain, not only the tag, so that separation can be evaluated as the reader would observe it at 45° and 150 mm.
- An improvement large enough to meet the recorded practical criterion.
- Fabrication and re-measurement of the candidate under the same four-position protocol, followed by a repeat of the strict evaluation of Section 4.2 on the new tag family.

Appendix G — Supporting Visualisations

G.1 Supporting representation and condition-block views

Figures 16 and 17 provide supporting views of results already discussed. Figure 16 uses the historical target-informed representation from Section 5.3, whereas Figure 17 uses the primary P1–P3 source-only model from Section 4.2; they answer different questions and should not be compared as model alternatives. The integrated three-level synthesis is presented in Table 27 (Section 6.1.2).

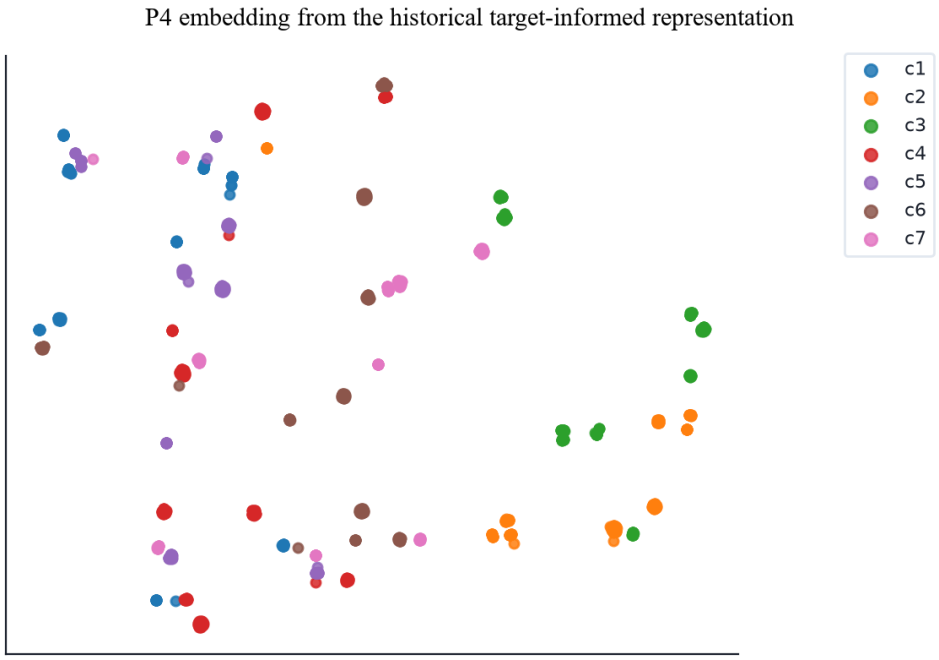

Figure 16. Two-dimensional t-SNE view of the P4 embedding from the historical target-informed representation. It is a supporting illustration only; no quantitative conclusion uses the projection [31]. Legend entries c1–c7 denote TagIDs 1–7. The display uses a fixed 900-row sample drawn without replacement from the 3,150 P4 measurement-row embeddings; each point represents one measurement row.

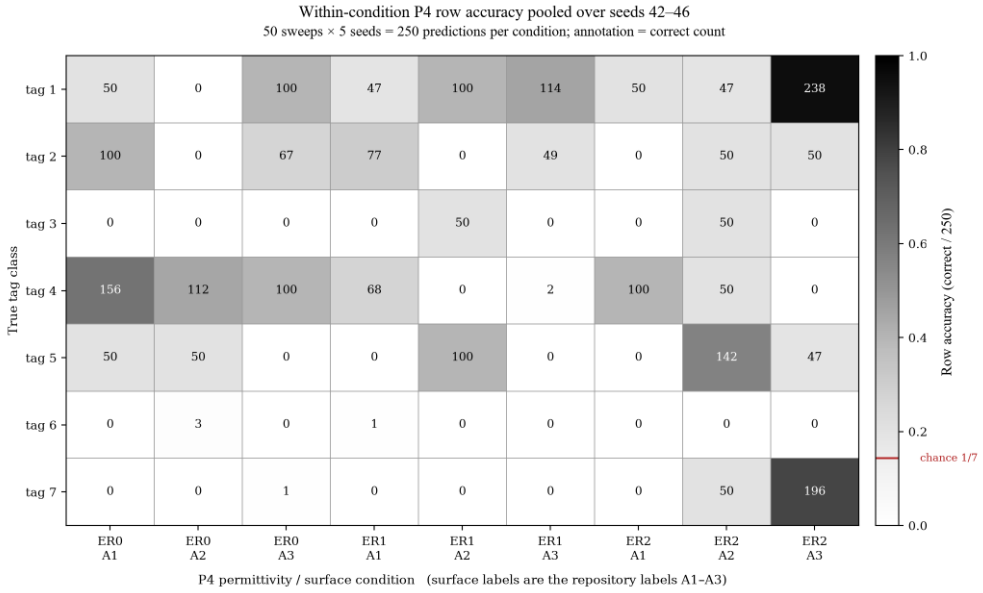

Figure 17. Within-condition P4 row accuracy pooled over seeds 42–46 for the primary source-only model. Each condition contributes 50 sweeps × 5 seeds = 250 predictions. Cell value = correct predictions / 250; printed annotations give the correct-prediction counts. This is a descriptive within-condition row statistic, not majority-vote condition-block Accuracy. Thirty of 63 conditions are never recovered; overall pooled row Accuracy remains 0.1566.

G.2 Supporting peak and frequency-domain diagnostics

Historical clarification indicates that the supplied 281 samples are a crop intended to span approximately 5–8 GHz, but the original 700-point frequency vector and exact crop provenance are not preserved. The four reported spectral regions therefore use a conditional linear endpoint mapping rather than recovered instrument coordinates.

The pre-specified ordered-index peak windows were W0 10–69, W1 75–126, W2 136–182 and W3 192–252. The descriptor did not pass its extraction-validity gate and is retained only as supporting associative evidence.

Table 35. Reference-peak regions under a conditional linear 281-point mapping to approximately 5–8 GHz. The original frequency vector and crop indices are not preserved, and peak evidence is associative only. Raw-response peaks and model-attribution regions are different quantities and are not expected to coincide exactly.

Feature	Raw response mean	Attribution plateau
Lc	5.340 GHz	5.299 – 5.346 GHz
L3	6.017 GHz	5.945 – 5.992 GHz
L2	6.685 GHz	6.650 – 6.697 GHz
L1	7.478 GHz	7.642 GHz

Figure 18 shows the calibration-peak behaviour. Four complementary diagnostics—raw-spectrum reconstruction, source-only occlusion, integrated-gradient attribution, and saved per-class and per-position tables—show frequency-localised structure in these regions. Raw-response peaks and model-attribution regions are different quantities and are not

expected to coincide exactly: Lc and L2 co-locate to within tens of megahertz, while L3 and L1 are offset. The diagnostics are therefore treated as consistent localisation, not as independent causal confirmation.

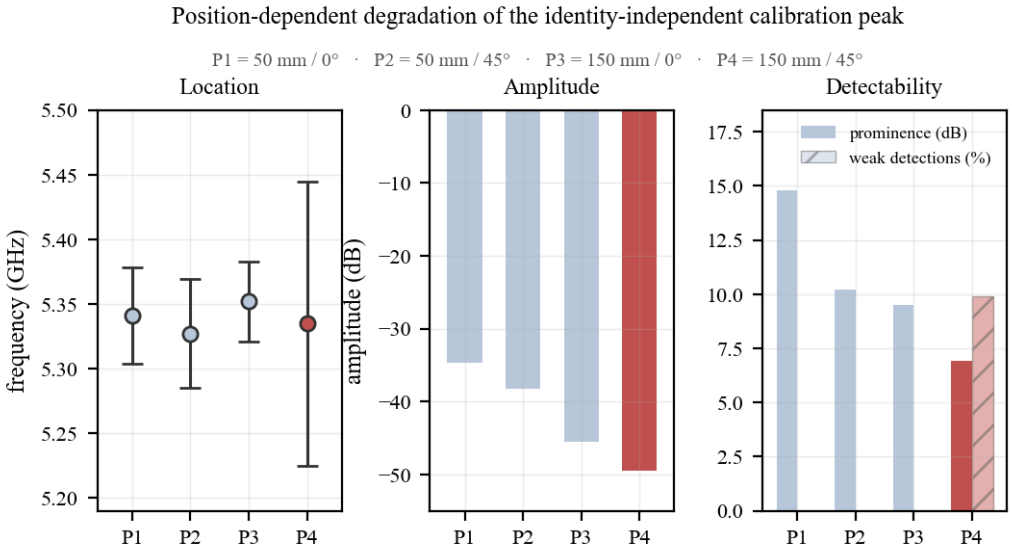

Figure 18. Position-dependent behaviour of the calibration peak Lc. The frequency mapping is conditional and the analysis is associative. Location: mean ± population SD over 3,150 spectra per position; weak detections indicate Lc prominence below 3 dB (see note).

Note. The population is the historical reference-peak diagnostic. Prominence is the detected Lc peak amplitude minus the median amplitude in its detection window; weak detections are the percentage of the 3,150 spectra at each position with prominence below 3 dB. This low-prominence criterion differs from the four-peak descriptor's missing-peak rate.

The global missing-peak rate of 0.117004 exceeded the pre-specified 0.05 gate, indicating fragility in the inherited window-based peak descriptor. Missingness was lowest at P4 (0.0400 versus 0.1387, 0.1529 and 0.1364 at P1–P3; P4 minus rest −0.1027, 95% [−0.1345, −0.0702]), so missing peaks do not explain the P4 transfer failure. The analysis supports associative localisation, not a complete electromagnetic causal explanation.

Source-domain peak separability is low in absolute terms: only 13 of 252 tag-pair × peak × position combinations (5.16%) are reliably separated in-domain, that is, separated by more than the within-tag scatter across the nine permittivity–surface conditions. That is a property of this four-peak descriptor rather than a direct measurement of physical codebook margin.