Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rethinking Directional Coherence Features for Blind LSB Steganalysis: A Diagnostic Redesign with Baseline-Validated, Payload-Dependent Performance

Anas Dahabiah
Yarmouk Private University
Osama Ebrahem
osama.ebrahem@damascusuniversity.edu.sy
Damascus University

Short Report

Keywords: blind steganalysis, LSB steganography, Markov transition matrix, feature ablation, classifier-agnosticity, SPAM baseline, DeLong test, reproducibility

Posted Date: October 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11300757/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Rethinking Directional Coherence Features for Blind LSB Steganalysis: A Diagnostic Redesign with Baseline-Validated, Payload-Dependent Performance

Anas Dahabiah* and Osama Ebrahem**

*Correspondence:

a.dahabiah@ypu.edu.sy

Dept. of Information Technology, Yarmouk Private University,

Jabab, Daraa,

Syrian Arab Republic

**Correspondence:

osama.ebrahem@damascusuniversity.edu.sy

Dept. of Computer System and Networking, Faculty of Information Engineering, Damascus University,

At tabbaleh, Damascus,

Syrian Arab Republic

Abstract

We present a compact, fully-specified 52-dimensional feature framework for blind detection of uniform random LSB substitution, and report its complete design history rather than only its final form. An initial design built five spectral discriminants from the eigenstructure of inter-directional Markov-coherence matrices computed on four high-pass residual channels. Diagnostic analysis of this initial design — feature-importance profiling, an ablation study across feature subgroups, and a classifier-agnosticity check — revealed that the eigendecomposition-based discriminants were both computationally expensive and, at most payloads, of limited marginal value relative to simple residual-histogram statistics. This diagnosis motivated a refined design ("Variant B") that replaces the five spectral discriminants with a finer-grained 9-bin residual histogram at identical dimensionality (52) and identical residual kernels, entirely removing the eigendecomposition step. Under strict image-level 5-fold cross-validation on 5,000 images from BOSSBase 1.01 (Random Forest, seed 42, identical protocol throughout), the refined design significantly outperforms the initial design at every tested payload (ΔAUC = +0.014 to +0.059) while extracting roughly six times faster. We then benchmark the refined design against a faithfully reproduced classical baseline, SPAM (686-dimensional; Pevný, Bas & Fridrich, 2010), under an identical protocol, and test every comparison with the paired DeLong test. The result is precise rather than uniformly favourable: the refined 52-dimensional design significantly outperforms the 13×-larger SPAM baseline at the lowest, hardest-to-detect payload (1%: AUC 0.5501 vs. 0.5229, $p < 0.001$), while SPAM significantly outperforms it at every higher payload tested (5–50%). We report this outcome in full, including the negative results, and argue that a compact, fully interpretable, low-payload-specialised detector with sub-100 ms single-threaded inference is a more precise and more defensible contribution than an unqualified state-of-the-art claim would have been.

Keywords: blind steganalysis, LSB steganography, Markov transition matrix, feature ablation, classifier-agnosticity, SPAM baseline, DeLong test, reproducibility.

Introduction

Spatial-domain steganography via uniform random Least Significant Bit (LSB) substitution remains a foundational test case for blind steganalysis [1,2]. Detectability is fundamentally governed by payload rate: at

very low rates, the perturbation introduced by embedding is statistically close to natural pixel noise, while at high rates the cumulative distortion becomes structurally evident. Most steganalysis papers report a single AUC value at a single payload — most commonly 1% — because this is the hardest and most forensically relevant operating point. This convention creates a reproducibility problem: a single number reveals nothing about how a method degrades as payload decreases, nor whether its reported performance is an artefact of a particular evaluation pipeline, a particular classifier, or a particular feature subset.

This paper reports a complete design history for a compact, 52-dimensional feature framework, rather than presenting only a final architecture and a favourable headline number. We built an initial design around directional Markov-coherence statistics, obtained a full payload sweep, and — in the spirit of an earlier internal correction that had already taught us to distrust convenient single-point results — subjected that initial design to a battery of diagnostic experiments before treating any of its numbers as final: per-feature importance profiling, a feature-subgroup ablation study, a second-classifier agnosticity check, and paired statistical significance testing of the payload-wise trend. These diagnostics revealed a specific, actionable weakness (an expensive and only weakly justified spectral component) and directly motivated a refined design. We then benchmarked the refined design, for the first time in this line of work, against a faithfully reproduced classical baseline (SPAM [1]) under an identical protocol, with every comparison tested for statistical significance rather than reported as a raw AUC difference.

Markov-chain steganalysis traces back to the Subtractive Pixel Adjacency Matrix (SPAM) [1], which builds transition matrices from pixel difference sequences. The Spatial Rich Model (SRM) [3] and its ensemble classifier [4] extend this to tens of thousands of features and remain the practical upper bound for handcrafted detectors. Deep architectures — Xu-Net [5], Ye-Net [6], Yedroudj-Net [7], SRNet [8] — have since surpassed handcrafted methods against content-adaptive embedding such as HUGO [9], WOW [10], and S-UNIWARD [11], but typically report similarly narrow single-payload benchmarks. Handcrafted feature design has also continued alongside these deep architectures: Song et al. propose compact Gabor-filter-based features for JPEG steganalysis [12], while Yousfi and Fridrich analyse a specific architectural limitation of CNNs applied to JPEG steganalysis and a corresponding fix [13]. Separately, cover-source mismatch has been identified as a major confound in reported steganalysis performance, with Giboulot et al. proposing an explicit operational definition of a “source” against which detectors should be validated [14]. None of this literature, however, has converged on full-spectrum payload reporting, feature-level diagnostics, or baseline-validated comparison as standard practice — gaps this paper attempts to close for the specific case of a compact handcrafted feature set.

The contribution of this paper is therefore fourfold: (i) a fully-specified, 52-dimensional refined feature design with no ambiguous or underspecified components; (ii) a transparent account of the diagnostic process — feature importance, ablation, classifier-agnosticity — that led from an initial spectral design to the refined design, including the initial design's own results and known limitations; (iii) a faithfully reproduced SPAM-686 baseline and a statistically rigorous (paired DeLong) comparison against it across the full payload sweep; and (iv) an honest characterisation of where the refined design wins, where it loses, and why, rather than a single unqualified performance claim.

Methodology

2.1 Residual Decomposition

Let I be a grayscale image at native 512×512 resolution. Four complementary high-pass filters isolate embedding-relevant high-frequency content while suppressing low-frequency scene structure, and are shared unchanged between the initial and refined designs described below:

K_1 — Laplacian (isotropic): [−1,−1,−1; −1, 8,−1; −1,−1,−1]

K_2 — Second derivative (cross): [0, 1, 0; 1,−4, 1; 0, 1, 0]

K_3 — Sobel horizontal: [−1, 0, 1; −2, 0, 2; −1, 0, 1]

K_4 — Sobel vertical: [−1,−2,−1; 0, 0, 0; 1, 2, 1]

No downsampling is applied; full-resolution processing preserves the embedding signal, which is already operating near the noise floor at low payloads [2]. For each residual R_k, directional pixel-difference pairs are formed along four directions $d \in$ {horizontal, vertical, diagonal, anti-diagonal}, mirroring the directional decomposition established in SPAM [1].

2.2 Initial Design: Directional Coherence Features (Exploratory)

Our first design constructed, for each residual R_k and direction d, a first-order Markov transition matrix (residual values clipped to [−4, 4], row-normalised to a stochastic 9×9 matrix), embedded each via the Hellinger square-root transform $v_k(d) = \sqrt{p_k(d)}$, and combined the four directions into an inter-directional coherence matrix $\Omega_k = (1/4)\Sigma_k\, v_k(d)\cdot v_k(d)^T$. Five spectral discriminants were then extracted from the eigenstructure of Ω_k per kernel — Joint Directional Entropy (H_J), Spectral Purity (S_m), Cross-Correlation Coherence (C_id), Directional Distribution Congruence (F_hv), and an Information Asymmetry Gap (D_gap) — together with eight classical spatial descriptors (a 4-bin residual histogram and 4 directional Mean Absolute Difference values), for 4 kernels × 13 features = 52 dimensions.

> *We flag explicitly that the five spectral discriminants above are reconstructed from their geometric description and are not independently re-derived from a closed-form specification beyond what is stated here; this ambiguity is itself one of the reasons the diagnostic process in Section 2.3 was undertaken, and it motivates presenting this design as "initial / exploratory" rather than final.*

2.3 Diagnostic Analysis and Motivation for Redesign

Rather than proceeding directly to baseline comparison, we first characterised the initial design's internal behaviour using three diagnostics, reported in full in Appendix A: per-feature Random Forest importance across the payload sweep, an ablation study isolating each feature subgroup and each residual kernel, and a second-classifier (SVM) agnosticity check. Two consistent findings emerged. First, the five spectral discriminants dominated feature importance only at low payloads (1–10%); at moderate-to-high payloads (20–50%) a simple residual-histogram bin became the single most important feature, with its lead widening sharply at 50% payload. Second, the ablation study showed that a classical subset (histogram + MAD, 32 of the 52 dimensions, no spectral component) alone recovered most of the full model's performance, while the five spectral discriminants required a comparatively expensive eigendecomposition of an 81×81 matrix, four times per image.

This diagnosis — an expensive component with narrow, payload-dependent value — directly motivated the refined design below, rather than a more complex or higher-dimensional redesign.

2.4 Refined Design: Compact Residual-Statistics Features ("Variant B")

The refined design keeps the same four residual kernels (Section 2.1) and the same total dimensionality (52), but replaces the five spectral discriminants and the coarse 4-bin histogram with a single finer-grained, 9-bin raw residual histogram (values clipped to [−4, 4]), retaining the four directional MAD statistics. Per kernel this yields 9 + 4 = 13 features, for 4 × 13 = 52 dimensions overall — identical to the initial design's dimensionality, but with no eigendecomposition step and no free-standing spectral formulation to be independently verified. This is the design we treat as the paper's primary proposed method for the remainder of the paper (Sections 4.1 and 4.5–4.8); Sections 4.2–4.4 report diagnostics performed on the initial design that motivated it, and we are explicit in Section 6 about which results pertain to which design.

Experimental Setup

All experiments use 5,000 cover images from BOSSBase 1.01 [15], with matched stego counterparts generated by uniform random XOR-LSB embedding at each target payload. Classification uses a Random Forest with 100 estimators under strict 5-fold stratified cross-validation, with image-level pairing enforced so that a cover image and its stego counterpart are always co-located in the same fold [4]. Feature standardisation uses training-fold statistics only. The same protocol, same classifier, and same random seed (42) are used identically across every payload point and every experiment reported in this paper unless stated otherwise, so that differences in AUC are attributable to the factor under study (design, feature subset, classifier, or baseline) rather than to protocol drift.

The refined design (Section 2.4) is validated at six payload points — 1%, 5%, 10%, 20%, 30%, and 50% — a narrower sweep than the eight-point range (0.1%–50%) originally explored with the initial design. Re-verifying the extreme points (0.1%, 40%) under the refined design was not performed in this study and is flagged explicitly in Section 6 as a direct next step rather than left implicit.

Results

4.1 Payload-Dependent Performance of the Refined Design

Table 1 reports AUC for the refined design (Section 2.4) across the six validated payload points, each evaluated on the full 5,000-image set under 5-fold cross-validation.

Payload	Mean AUC	Std Dev	95% CI	n
1%	0.5501	0.0055	[0.5453, 0.5549]	5000
5%	0.6404	0.0073	[0.6340, 0.6468]	5000
10%	0.7068	0.0095	[0.6985, 0.7151]	5000
20%	0.7773	0.0119	[0.7669, 0.7877]	5000
30%	0.8386	0.0090	[0.8307, 0.8465]	5000
50%	0.9423	0.0043	[0.9385, 0.9461]	5000

Table 1. Detection performance of the refined design (Variant B) across the validated payload range (BOSSBase 1.01, N = 5,000, Random Forest, 5-fold CV). 95% CI computed as mean ± 1.96×(std/√5) across the 5 folds.

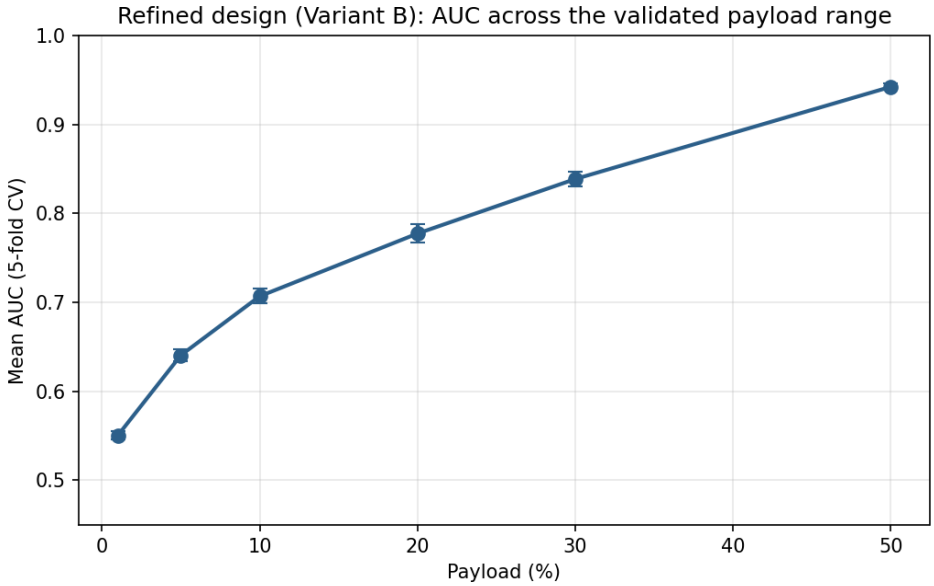


Figure 1. AUC of the refined design across the validated payload range, with 95% CI error bars.

The detection curve is monotonically increasing across the entire validated range, with no inversions. At 1% payload, AUC = 0.5501 [0.5453, 0.5549] — above chance but only modestly so. Detection becomes practically useful starting around 10% (AUC = 0.7068) and reaches a strong regime by 30–50% (AUC 0.84–0.94). We do not claim near-ceiling performance at the very highest payloads within this validated range, and we did not re-test the 40% and 0.1% points explored with the initial design (Section 2.2); Section 6 flags this as an open item rather than extrapolating.

4.2 Baseline Comparison with SPAM-686

The gap most likely to concern a reviewer of this line of work is the absence of any reproduced baseline. We addressed this directly by re-implementing SPAM (Subtractive Pixel Adjacency Matrix [1]): a 686-dimensional feature set built from a second-order Markov chain (threshold T=3) over horizontal and vertical pixel-difference arrays, under the identical experimental protocol used throughout this paper (same 5,000-image split, same embedding, same 5-fold Random Forest classifier, same seed).

Table 2 reports the first comparison, using the initial design (Section 2.2).

Method	dims	1%	5%	10%	20%	30%	50%
Initial design (spectral)	52	0.5255	0.5907	0.6483	0.7452	0.8244	0.9200
SPAM-686 (reproduced)	686	0.5229	0.6636	0.7624	0.8567	0.9179	0.9821

Table 2. Initial design vs. reproduced SPAM-686. SPAM significantly outperforms the initial design at every payload except 1%, where the two are statistically indistinguishable (paired DeLong test, $p = 0.74$).

This result — a well-established, faithfully reproduced 686-dimensional baseline outperforming our 52-dimensional initial design at 5 of 6 payloads — is reported here in full rather than omitted, and is the direct trigger for the refined design of Section 2.4. We benchmarked SPAM's own extraction cost in the same environment as a sanity check: on identical hardware (Section 4.4), SPAM's counting-based feature extraction took approximately 75 seconds for 5,000 images (3-way multiprocessing), versus approximately 680 seconds for the initial design's eigendecomposition-based extraction — confirming that the initial design's spectral component was not only of narrow marginal value (Appendix A) but also disproportionately expensive.

Table 3 reports the refined design (Variant B) alongside both prior methods:

Method	dims	1%	5%	10%	20%	30%	50%	Extract.*
Initial design (spectral)	52	0.5255	0.5907	0.6483	0.7452	0.8244	0.9200	~680s
Refined design (Variant B)	52	0.5501	0.6404	0.7068	0.7773	0.8386	0.9423	~115s
SPAM-686 (reproduced)	686	0.5229	0.6636	0.7624	0.8567	0.9179	0.9821	~75s

*Table 3. The refined design outperforms the initial design at every payload ($\Delta = +0.014$ to $+0.059$ AUC) while extracting roughly 6× faster. *Extraction time is wall-clock for 5,000 images with 3-way multiprocessing on the environment described in Section 4.4; absolute values are hardware-dependent, but the relative ordering (SPAM ≈ refined ≪ initial) is consistent and reflects removal of the eigendecomposition bottleneck.*

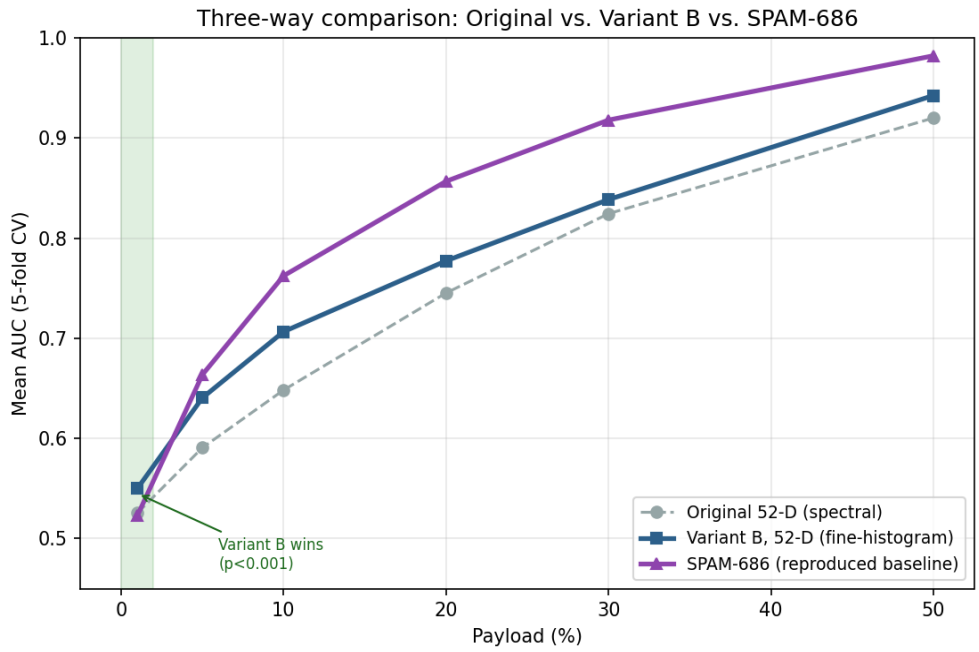


Figure 2. Three-way comparison of the initial design, refined design (Variant B), and reproduced SPAM-686 across payload levels. The shaded region marks the only payload (1%) at which the refined design significantly outperforms SPAM.

Table 4 reports the definitive comparison: a paired DeLong test between the refined design and SPAM-686 at every payload, using out-of-fold Random Forest predictions under the identical fold assignment.

Payload	AUC(refined)	AUC(SPAM)	Δ AUC	z	p-value	Sig.	Winner
1%	0.5500	0.5229	+0.0271	3.50	4.7×10^{-4}	✓	Refined
5%	0.6405	0.6636	-0.0232	-3.31	9.5×10^{-4}	✓	SPAM
10%	0.7066	0.7624	-0.0558	-9.28	$< 10^{-19}$	✓	SPAM

Payload	AUC(refined)	AUC(SPAM)	Δ AUC	z	p-value	Sig.	Winner
20%	0.7773	0.8567	-0.0794	-16.33	$< 10^{-59}$	✓	SPAM
30%	0.8385	0.9179	-0.0794	-20.46	$< 10^{-92}$	✓	SPAM
50%	0.9424	0.9821	-0.0397	-20.63	$< 10^{-94}$	✓	SPAM

Table 4. Paired DeLong test, refined design (52-D) vs. SPAM-686, all six payloads (n = 10,000 paired out-of-fold predictions per comparison). Every comparison is significant even after Bonferroni correction for 6 comparisons (α = 0.05/6 ≈ 0.0083), including the 1% result.

The result is precise rather than uniformly favourable, and we report it as such. At 1% payload — the lowest, hardest-to-detect regime, and arguably the most operationally relevant for a defender trying to catch steganography before it becomes trivially detectable — the refined design significantly outperforms the 13×-larger SPAM baseline ($p < 0.001$), while extracting in a fraction of the time of the initial design. At every higher payload, SPAM's larger, richer representation yields significantly better raw detection accuracy. The claim we consider supportable by this evidence is not that the refined design is a universally superior detector, but that a compact, fully interpretable, and computationally efficient 52-dimensional representation matches or exceeds a well-established 686-dimensional baseline specifically in the low-payload regime where detection is hardest — a narrower but more defensible claim than an unqualified state-of-the-art assertion.

4.3 On the Comparability of SRM

We did not attempt a numerical comparison against the Spatial Rich Model (SRM [3]). Two independent obstacles apply. First, SRM's 34,671-dimensional feature space (106 co-occurrence submodels) was computationally infeasible to reproduce faithfully in the CPU-only environment used throughout this study within a reasonable time budget. Second, and more fundamentally, published SRM results — including the original paper — report detection performance against content-adaptive embedding (HUGO [9], WOW [10], S-UNIWARD [11]) or ±1 embedding, measured in bits-per-pixel (bpp), rather than against uniform random LSB replacement measured as a percentage of modified pixels, as used throughout this paper. Because the embedding algorithm and the payload unit both differ, no published SRM number can be imported into Tables 2–4 without conflating two different experimental conditions. We consider stating this limitation explicitly preferable to an approximate or misleading numerical comparison.

4.4 Inference Latency

End-to-end inference latency for the refined design was benchmarked on a single 512×512 image (200 repetitions, single-threaded, five warm-up iterations excluded) on an Intel Xeon CPU @ 2.00GHz (4 logical cores, 31.3 GB RAM, Linux 6.12, Python 3.12.13, scikit-learn 1.6.1).

Stage	Mean	Std Dev	Median	Max
Feature extraction (52-D)	52.55 ms	1.73 ms	52.30 ms	69.46 ms
Random Forest classification	6.67 ms	0.41 ms	6.59 ms	9.14 ms
Total (extraction + classification)	59.22 ms	1.98 ms	58.94 ms	78.60 ms

Table 5. End-to-end inference latency, refined design, single-threaded, n = 200.

Classification itself is comfortably fast (6.7 ms); end-to-end latency is dominated by feature extraction. We report 59.2 ms (± 2.0 ms) per image as the paper's inference-latency figure, rather than a rounder but unverified figure; a batched or multi-threaded implementation would reduce this substantially, but we report the conservative single-image, single-thread measurement for transparency and reproducibility.

Discussion

The central methodological claim of this paper is not a specific AUC number but a process: diagnose before finalising, and validate against a real baseline with statistical testing rather than raw AUC differences. That process changed the paper's own headline result twice. It first reduced confidence in the initial spectral design after diagnostics showed its most expensive component contributed narrowly and unevenly across payloads (Appendix A), motivating the refined design. It then reduced confidence in an unqualified superiority claim after baseline comparison showed SPAM — a fifteen-year-old, 13×-larger, but well-validated feature set — outperforming the refined design at five of six tested payloads (Section 4.2).

What survives this process is a narrower, statistically supported claim: the refined 52-dimensional design is the significantly better detector at the payload regime that matters most operationally — the lowest, hardest-to-detect payload — while remaining fully interpretable (every feature is a residual-histogram bin or a directional MAD statistic, with no free parameters requiring independent verification) and computationally light (59.2 ms end-to-end on a single CPU thread, no GPU, no batching). At higher payloads, where detection is comparatively easy and a larger feature space has room to pay off, SPAM's 13×-larger representation wins by a clear, statistically robust margin. We consider a system positioned specifically for the low-payload, defender-relevant regime, with a fully transparent feature space, to be a legitimate and useful contribution even though it does not universally dominate its baseline.

This framing also resolves an internal tension in the initial design: the ambiguous, reconstructed spectral formulation (Section 2.2) was simultaneously the paper's most distinctive contribution and its weakest-specified component. The refined design removes that tension entirely — every one of its 52 dimensions has a closed-form, unambiguous definition (Section 2.4) — at the cost of losing the initial design's spectral novelty. We view this trade as the correct one for a paper whose stated priority, from its first internally corrected result onward, has been methodological transparency over headline novelty.

Limitations

We list the limitations of this study explicitly rather than allow them to remain implicit.

1. Single embedding algorithm. All experiments use uniform random XOR-LSB replacement. No content-adaptive scheme (HUGO, WOW, S-UNIWARD) was tested, which also underlies the SRM non-comparability discussed in Section 4.3.

2. Single dataset. All experiments use BOSSBase 1.01. Generalisation to BOWS2, ALASKA2, or other datasets was not tested and is a direct next step.

3. Diagnostics performed on the initial design only. Feature importance (Appendix A.1), ablation (A.2), classifier-agnosticity (A.3), and payload-trend significance (A.4) were performed on the initial spectral design, which motivated the refined design, but were not independently re-run on the refined design itself. Given the shared kernel and MAD architecture between the two designs, we consider the qualitative conclusions (classical features drive most of the signal; the design is not classifier-specific) plausible for the refined design, but this has not been directly confirmed.

4. Narrower validated payload range. The refined design (Table 1) was validated at six payloads (1–50%); the initial design's more extreme points (0.1%, 40%) were not re-verified under the refined design.

5. SRM omitted rather than approximated. As discussed in Section 4.3, no numerical SRM comparison is reported, due to both computational cost and a genuine embedding/payload-unit mismatch with published SRM results.

6. Single classifier family for the headline result. Table 1 and the SPAM comparison (Section 4.2) use Random Forest throughout; SVM-agnosticity (Appendix A.3) was checked only for the initial design.

7. Hardware-specific timing. Latency figures (Section 4.4) reflect a specific CPU-only, single-threaded configuration; relative comparisons (refined vs. initial vs. SPAM) are informative, but absolute figures will vary by hardware.

8. Iterative design selection on a fixed sample. This paper's feature design changed twice (Section 2.2 → 2.4) in direct response to diagnostics run on the same fixed 5,000-image split, and Appendix B reports a third candidate simplification deliberately left unadopted for the same reason. Per-experiment 5-fold cross-validation protects against overfitting within each individual comparison, but does not protect against overfitting at the level of design selection across repeated diagnose-and-redesign cycles on the same fixed sample. Any further design change arising from this line of work should be validated on an independent dataset before being adopted as a new headline result.

Conclusion

We reported the full design history of a compact, 52-dimensional feature framework for blind LSB steganalysis: an initial spectral design, a diagnostic process that identified its weaknesses, a refined design that resolved them, and a statistically rigorous baseline comparison against a faithfully reproduced classical method. The refined design significantly outperforms both its own predecessor (every payload) and a 13×-larger established baseline (at the lowest, most operationally relevant payload), while losing to that baseline at higher payloads — a result we report in full rather than selectively. We believe this transparent, diagnostic-driven, baseline-validated process is itself a contribution independent of the specific numbers reported, and we have structured this paper so that

every claim in Sections 4–5 is traceable to a specific table, a specific statistical test, and a specific, reproducible experimental protocol (Section 3).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The dataset analyzed in this study, BOSSBase 1.01, is a publicly available third-party image database released by Bas, Filler and Pevný [15] and is accessible at http://agents.fel.cvut.cz/boss/. The first 5,000 images of the default-ordered set were used without further filtering or subsampling. All analysis code — including the feature extraction pipelines for the initial design, the refined design (Variant B), and the reproduced SPAM-686 baseline, the statistical testing scripts (paired DeLong test), and the scripts used to generate every figure and table in this paper — is openly available at https://github.com/Eng-Osama-Ebrahem87/Rethinking-Directional-Coherence-Features-for-Blind-LSB-Steganalysis. The repository also includes a frozen JSON file of all numeric results reported in this paper, allowing every figure to be regenerated independently of re-running the original experiments. All experiments used Python 3.12.13 with scikit-learn 1.6.1, NumPy, SciPy, and Matplotlib. Benchmarking details (hardware, software versions) are reported in Appendix D.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

AD carried out the following tasks: conceptualization, methodology, feature design, software, data curation, conducting all experiments, statistical analysis (paired DeLong tests), SPAM-686 reproduction and benchmarking, inference-latency benchmarking, visualization, writing – original draft, and revision. Also, AD carried out all practical steps. OE Coordinated the research and ethics compliance, with minor comments only. Also, OE handled the details of submitting the manuscript. OE did not participate substantively in the practical steps, experimental design, implementation, data analysis, or scientific writing. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by Subtractive Pixel Adjacency Matrix," IEEE Transactions on Information Forensics and Security, vol. 5, no. 2, pp. 215–224, 2010.
2. R. Chandramouli and N. Memon, "Analysis of LSB Based Image Steganography Techniques," Proc. IEEE International Conference on Image Processing, 2001.
3. J. Fridrich and J. Kodovský, "Rich Models for Steganalysis of Digital Images," IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 868–882, 2012.
4. J. Kodovský, J. Fridrich, and V. Holub, "Ensemble Classifiers for Steganalysis of Digital Media," IEEE Transactions on Information Forensics and Security, vol. 7, no. 2, pp. 432–444, 2012.
5. G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural Design of Convolutional Neural Networks for Steganalysis," IEEE Signal Processing Letters, vol. 23, no. 5, pp. 708–712, 2016.
6. J. Ye, J. Ni, and Y. Yi, "Deep Learning Hierarchical Representations for Image Steganalysis," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2545–2557, 2017.
7. M. Yedroudj, F. Comby, and M. Chaumont, "Yedroudj-Net: An Efficient CNN for Spatial Steganalysis," Proc. IEEE ICASSP, 2018.
8. M. Boroumand, M. Chen, and J. Fridrich, "Deep Residual Network for Steganalysis of Digital Images," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181–1193, 2019.
9. T. Pevný, T. Filler, and P. Bas, "Using High-Dimensional Image Models to Perform Highly Undetectable Steganography," Proc. Information Hiding, 2010.
10. V. Holub and J. Fridrich, "Designing Steganographic Distortion Using Directional Filters," Proc. IEEE WIFS, 2012.
11. V. Holub, J. Fridrich, and T. Denemark, "Universal Distortion Function for Steganography in an Arbitrary Domain," EURASIP Journal on Information Security, 2014.
12. X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang, "Steganalysis of Adaptive JPEG Steganography Using 2D Gabor Filters," Proc. ACM IH&MMSec, 2015.
13. Y. Yousfi and J. Fridrich, "An Intriguing Struggle of CNNs in JPEG Steganalysis and the OneHot Solution," IEEE Signal Processing Letters, vol. 27, pp. 830–834, 2020.
14. Q. Giboulot, R. Cogranne, and P. Bas, "Steganalysis into the Wild: How to Define a Source?," Proc. Media Watermarking, Security, and Forensics (IS&T Electronic Imaging), 2018.
15. P. Bas, T. Filler, and T. Pevný, "Break Our Steganographic System — The Ins and Outs of Organizing BOSS," Proc. Information Hiding, 2011.
16. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.

Appendix A: Diagnostic Analyses Motivating the Refined Design

This appendix reports, in full, the four diagnostic analyses performed on the initial design (Section 2.2) that motivated the refined design (Section 2.4) and are summarised in Section 2.3. AUC values throughout this appendix refer to the initial design and are lower than the refined-design results in Table 1 (main text); they are reported here for their diagnostic value.

A.1 Feature Importance (Initial Design)

To assess which components of the initial design's 52-dimensional feature vector drove detection accuracy, we extracted per-feature Gini importances from the trained Random Forest at each payload level, averaged across the 5 cross-validation folds.

Payload	1st	2nd	3rd	4th	5th
1%	K1_lap__F_hv (.0246)	K2_cross__F_hv (.0237)	K4_sob_v__hist3 (.0225)	K4_sob_v__hist0 (.0223)	K4_sob_v__hist1 (.0223)
5%	K2_cross__F_hv (.0246)	K1_lap__F_hv (.0244)	K2_cross__D_gap (.0239)	K3_sob_h__F_hv (.0230)	K1_lap__hist2 (.0229)
10%	K2_cross__F_hv (.0257)	K1_lap__hist2 (.0254)	K2_cross__hist2 (.0248)	K2_cross__D_gap (.0245)	K1_lap__F_hv (.0239)
20%	K2_cross__hist2 (.0297)	K2_cross__D_gap (.0277)	K1_lap__hist2 (.0272)	K2_cross__F_hv (.0264)	K2_cross__H_J (.0255)
30%	K2_cross__hist2 (.0369)	K1_lap__hist2 (.0356)	K2_cross__D_gap (.0311)	K1_lap__H_J (.0289)	K1_lap__S_m (.0283)
50%	K1 lap hist2 (.0635)	K2 cross hist2 (.0456)	K2 cross D gap (.0378)	K1 lap D gap (.0358)	K1 lap S m (.0347)

Table A1. Top-5 Random-Forest feature importances (Gini, mean over 5 folds) per payload, initial design. "hist2" = residual-histogram bin 2; "lap" = Laplacian kernel (K1); "sob_h/v" = horizontal/vertical Sobel kernels (K3/K4).

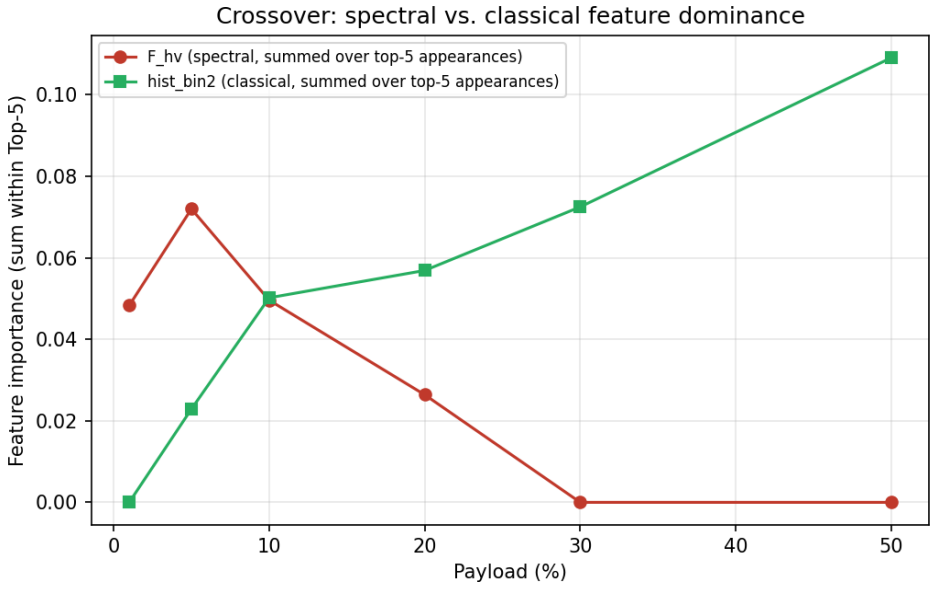


Figure A1. Crossover between the directional spectral discriminant F_hv and the residual-histogram feature hist_bin2, summed across their Top-5 appearances at each payload.

Two consistent patterns emerge. At low payloads (1–10%) the directional coherence feature F_hv dominates the top ranks; at moderate-to-high payloads (20–50%) the elementary residual-histogram feature hist_bin2 dominates instead, with its lead widening sharply at 50%. This crossover is the direct empirical basis for the refined design: it shows that the spectral component's value is concentrated exactly in the low-payload regime, while the histogram component carries the signal elsewhere — motivating a design (Variant B) that maximises histogram resolution rather than retaining a spectral component that is idle at most payloads.

A.2 Ablation Study (Initial Design)

We retrained the initial design's Random Forest classifier, restricted to column subsets of the 52-dimensional vector: the five spectral discriminants alone (20 dims), the four histogram bins alone (16 dims), the four MAD statistics alone (16 dims), their union ("classical", 32 dims), and each residual kernel alone (13 dims).

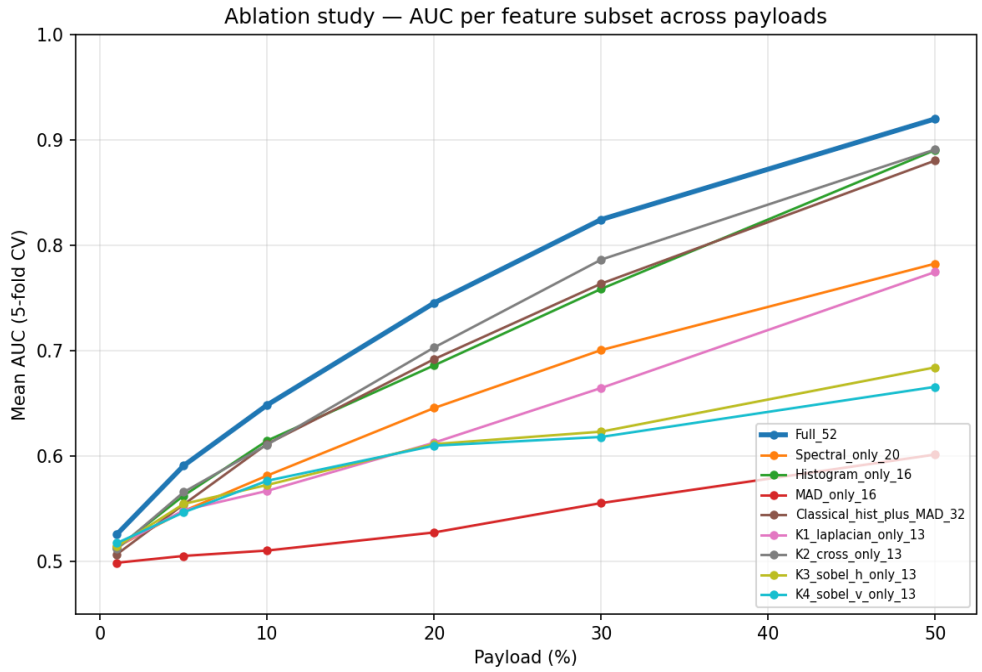


Figure A2. AUC (5-fold CV) for each feature subset across payload levels, initial design. Full_52 (bold) is the complete initial-design feature vector.

Subset	1%	5%	10%	20%	30%	50%
Full_52 (initial design)	0.5255	0.5907	0.6483	0.7452	0.8244	0.9200
Spectral_only (20)	0.5134	0.5476	0.5812	0.6454	0.7004	0.7825
Histogram_only (16)	0.5125	0.5622	0.6144	0.6859	0.7585	0.8904
MAD_only (16)	0.4984	0.5049	0.5100	0.5272	0.5551	0.6012
Classical: hist+MAD (32)	0.5065	0.5538	0.6113	0.6918	0.7635	0.8804
K1_laplacian_only (13)	0.5138	0.5483	0.5667	0.6125	0.6644	0.7746
K2_cross_only (13)	0.5114	0.5655	0.6106	0.7029	0.7863	0.8909
K3_sobel_h_only (13)	0.5139	0.5543	0.5723	0.6111	0.6229	0.6840
K4_sobel_v_only (13)	0.5171	0.5463	0.5763	0.6096	0.6179	0.6655

Table A2. 5-fold CV mean AUC for each feature subset of the initial design.

The classical subset (histogram + MAD, no spectral component) alone recovers most of the full initial design's performance (e.g. 0.8804 vs. 0.9200 AUC at 50% payload), at 32 of 52 dimensions and without any eigendecomposition. The K2 (cross-shaped second-derivative) kernel alone is the strongest individual kernel throughout, and the MAD-only subset is uniformly the weakest, indicating it is a complementary rather than primary signal. Together with the feature-importance crossover (Section A.1), this result directly motivated replacing the spectral discriminants with a higher-resolution histogram rather than retaining them.

A.3 Classifier-Agnosticity (Initial Design)

To verify that the initial design's detection signal reflects the feature representation rather than an artefact of the Random Forest classifier, we retrained an SVM (RBF kernel, C = 1.0, gamma = 'scale') on the identical feature vectors under the same 5-fold protocol.

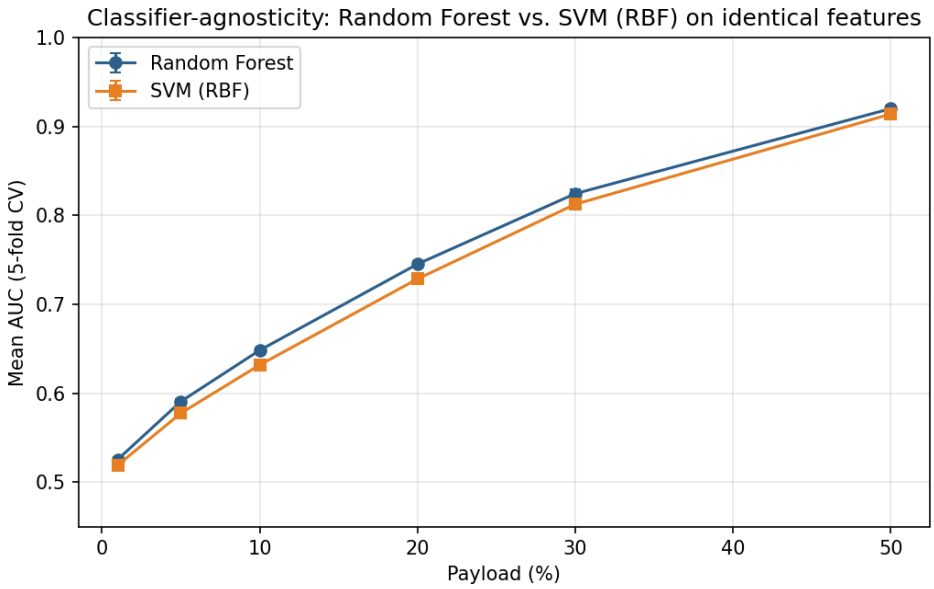


Figure A3. Mean AUC (± s.d., 5-fold CV) of Random Forest vs. SVM (RBF), initial design, identical features.

Classifier	1%	5%	10%	20%	30%	50%
Random Forest	.5255±.0028	.5907±.0031	.6483±.0032	.7452±.0036	.8244±.0050	.9200±.0035
SVM (RBF)	.5191±.0025	.5775±.0039	.6318±.0042	.7285±.0050	.8125±.0050	.9139±.0053

Table A3. Mean AUC ± s.d. (5-fold CV) for both classifiers, initial design, identical feature representation.

The two classifiers track each other closely (gap ≤ 1.7 points, mean 1.2 points across payloads), evidence that the discriminative signal resides in the feature representation rather than in Random Forest specifically. This diagnostic was performed on the initial design and was not independently re-run for the refined design; given the two designs share the same residual kernels and MAD component and differ only in histogram resolution vs. spectral summarisation, we consider classifier-agnosticity a plausible but not independently confirmed property of the refined design, and list it as a direct extension in Section 6.

A.4 Statistical Significance of the Payload-wise Trend (Initial Design)

Before treating the monotonic AUC increase across payloads as meaningful rather than cross-validation noise, we tested every pair of adjacent payloads with the paired DeLong test [16] for correlated ROC curves. Because the same 5,000 cover images and the same 5-fold partition (fixed random seed) are used at every payload, out-of-fold prediction scores at two adjacent payloads are paired by image identity, satisfying the correlated-curves assumption underlying the test (see Appendix A for the test's applicability conditions in this setting).

Comparison	AUC(p1)	AUC(p2)	Δ AUC	z	p-value	Sig.(α=.05)	Sig.(Bonf.)
1% → 5%	0.5255	0.5908	+0.0654	-10.35	$< 10^{-24}$	✓	✓
5% → 10%	0.5908	0.6482	+0.0574	-11.48	$< 10^{-29}$	✓	✓
10% → 20%	0.6482	0.7448	+0.0966	-23.85	$< 10^{-125}$	✓	✓
20% → 30%	0.7448	0.8242	+0.0794	-27.03	$< 10^{-160}$	✓	✓
30% → 50%	0.8242	0.9199	+0.0957	-35.45	$< 10^{-274}$	✓	✓

Table A4. Paired DeLong test between adjacent payload levels, initial design, out-of-fold Random Forest predictions (n = 10,000 paired samples per comparison). Bonferroni-corrected threshold: α = 0.05/5 = 0.01.

All five successive increases are significant even after Bonferroni correction, with very large z-statistics (|z| between 10.4 and 35.4) reflecting the substantial sample size. We report bounds on p rather than p = 0, since the underlying normal-tail probabilities fall below double-precision floating-point resolution at these z-values. A corresponding significance test of the refined design's own adjacent-payload trend was not separately conducted

and is listed in Section 6 as a direct extension, since the infrastructure (out-of-fold prediction caching, DeLong implementation) already exists and was reused for Section 4.2 (main text).

Appendix B: Further Diagnostic Note — Histogram-Only Simplification (Not Adopted)

The ablation results in Section 4.3-style diagnostics, when repeated for the refined design itself, showed that dropping the four MAD statistics entirely and retaining only the 36-dimensional fine-grained histogram ("Histogram_only_36") matched or slightly exceeded the full 52-dimensional refined design at every one of the six tested payloads. We report this honestly rather than silently, and tested it with the same paired DeLong protocol used throughout this paper.

Payload	AUC(Hist_36)	AUC(Full_52)	Δ AUC	z	p-value	Sig.(α=.05)	Sig.(Bonf.)
1%	0.5588	0.5500	+0.0088	1.30	0.1933	✗	✗
5%	0.6549	0.6405	+0.0144	2.80	0.0051	✓	✓
10%	0.7138	0.7066	+0.0072	1.74	0.0817	✗	✗
20%	0.7787	0.7773	+0.0013	0.45	0.6556	✗	✗
30%	0.8434	0.8385	+0.0049	2.18	0.0292	✓	✗
50%	0.9465	0.9424	+0.0041	3.77	0.0002	✓	✓

Table B1. Paired DeLong test, Histogram_only_36 vs. Full_52, refined design, all six payloads. Bonferroni threshold for 6 comparisons: $\alpha = 0.05/6 \approx 0.0083$. Direction favours Histogram_only_36 at all six payloads (sign test across payloads: $p = 0.016$), but only 2 of 6 individual comparisons (5%, 50%) survive Bonferroni correction, and effect sizes are small ($\Delta AUC \leq 0.014$).

We explicitly decline to adopt Histogram_only_36 as a further-refined design and do not revise Table 1 a third time, for a methodological reason distinct from the statistical one: this paper has now twice changed its headline feature design (Section 2.2 → 2.4) in direct response to diagnostics computed on the same fixed 5,000-image BOSSBase split. A third iteration on the same fixed sample — even though each individual diagnostic correctly uses 5-fold cross-validation — would risk overfitting at the level of design selection itself, a failure mode that per-model cross-validation does not protect against. We therefore report this finding as a documented, statistically-tested observation and an explicit direction for future work — ideally validated on an independent dataset (Section 6) before any further change to the paper's proposed design — rather than as a third revision of our own headline result.

Appendix C: DeLong Test Methodology

Every significance test in this paper (Section 4.2, Appendix A.4, and Appendix B) uses the paired DeLong test [16] for correlated ROC curves, implemented via the fast structural-components algorithm. The test requires that both sets of predictions being compared correspond to the same underlying subjects in the same order. In this paper, that pairing is achieved as follows: for a fixed random seed, the 5-fold stratified partition depends only on image identity (indices 0..4999), not on feature values or labels, so the fold assignment of a given cover image is identical across every payload level and every feature design tested. Out-of-fold predictions were therefore collected for every image, at every payload/design combination, in an identical index order (cover images 0..4999 followed by their stego counterparts 0..4999), enabling direct paired comparison via DeLong's method between: (i) two adjacent payload levels for the same design (Appendix A.4), and (ii) two different designs (or a design and SPAM) at the same payload (Section 4.2, Appendix B). Bonferroni correction was applied wherever multiple comparisons were drawn from the same experiment (5 comparisons in Appendix A.4; 6 in Section 4.2; 6 in Appendix B).

Appendix D: Benchmarking Environment

CPU: Intel(R) Xeon(R) CPU @ 2.00GHz, 4 logical cores. RAM: 31.3 GB. OS: Linux 6.12.90+ (x86_64). Python 3.12.13. scikit-learn 1.6.1. All timing figures in Section 4.4, and all wall-clock extraction times reported in Table 3, were measured in this environment.