Verity

A Calibrated, Auditable Deployment Layer for Forensic Surface Comparison

Eric Hare
ericrhare@gmail.com

June 2026 | Technical White Paper

Abstract

Two decades of community review — the National Research Council's 2009 report and the 2016 PCAST report — found that the feature-comparison forensic disciplines, firearms and toolmark identification among them, lack characterized error rates. Cuellar et al. (2024) sharpened the point: *no* black-box study of firearm comparison has yet properly characterized one. Verity does not claim to supply the missing universal error rate; we argue no single number exists. Instead it operationalizes the response a statistician would give: report a *calibrated weight of evidence* — a likelihood ratio (LR) — with a *characterized cost* (C_{llr}) on a *named* reference population, and refuse to emit an LR outside the domain on which it was validated.

Verity is, deliberately, not a new matching algorithm. It is the production, deployment, and calibration layer for the CSAFE / Iowa State research program (Hare, Hofmann & Carriquiry 2017) and the NIST congruent-matching line (Song 2013; Song et al. 2018). Its contributions are five: (1) *unification* — one calibrated pipeline that scores striated, impressed, and fractured marks through *Congruent Matching Regions* (CMR), a dimension-agnostic generalization of Song's Congruent Matching Cells; (2) a calibrated-LR-with-honest-scope answer to Cuellar et al.; (3) a glass-box decision layer — region-level attribution plus a monotone, bounded calibration we call an *audit firewall*; (4) production engineering — a native Rust X3P codec, language bindings, and a deployed calibrated-LR API; and (5) a small methodological result — a `diag_contrast` scorer, selected over the Phase-1 mean-diagonal score and over multivariate fusion *by data, not assumption*. On the Hamby-252 bullet benchmark Verity attains a barrel-disjoint (source-disjoint) test $C_{\text{llr}} = 0.113 \pm 0.066$ with test AUC 1.000; we report validation results per study and never pool across firearm makes. Beyond bullets, the *same* pipeline under *one* scorer configuration is validated source-disjoint on impressed cartridge breech faces (Fadul) and striated screwdriver toolmarks (`tmaRks`), where CMR instantiates the counting principles of Song's Congruent Matching Cells and of the Chumbley statistic and approaches each specialist's performance — one calibrated algorithm, three reductions, each *measured* rather than asserted.

1 Introduction: a discipline without a characterized error rate

Firearm and toolmark identification asks whether two marked surfaces — two bullets, two cartridge cases, two toolmarks — were produced by the same source. For most of its history the answer has rested on a trained examiner's holistic judgement of "sufficient agreement." The 2009 National Research Council report *Strengthening Forensic Science in the United States* [5] found that the pattern-comparison disciplines had not been shown, with the rigor of measurement science, to do what they claim. The 2016 report of the President's Council of Advisors on Science and Technology (PCAST) [6] made the requirement explicit: a feature-comparison method is *foundationally valid* only if it has been empirically tested on appropriate samples and has a *characterized error rate*.

Cuellar, Vanderplas, Luby & Rosenblum (2024) [7] examined the black-box studies offered in response and concluded that *every one* of them has methodological problems severe enough that none establishes an error rate one could responsibly carry into court: inappropriate sampling, set-based designs that leak class information, inconclusives handled in ways that flatter the

method, and — most relevant here — error rates reported as if they were properties of the *discipline* rather than of a particular study population.

Verity takes a deliberately narrower and, we argue, more honest stance. There is no single "error rate of firearms identification," because the answer depends on the firearm population, the manufacturing process, the wear state, and the scan quality. What *can* be reported, defensibly, is:

1. a **likelihood ratio** — how much more probable the observed surface agreement is if the two marks share a source than if they do not — which is prior-independent and is the form of evidence the courts and the forensic-statistics literature ask for;
2. the **calibration cost** C_{llr} [13] of the system that produced that LR, measured *on a named reference population* under a *source-disjoint* protocol, so the number means "how well this system is calibrated on data like this," not "how often examiners are wrong";
3. an explicit **scope**: outside the validated domain, Verity declines to emit an LR rather than extrapolate.

This is the same move that forensic speaker comparison made years ago, and it is why Verity borrows that field's evaluation machinery. The remainder of this paper describes the lineage Verity builds on (§2), the comparison method and its cross-modality generalization (§3), the calibration firewall that turns a score into evidence (§4), the source-disjoint validation that answers Cuellar et al. directly (§5), the limitations we are candid about (§6), and the open platform on which all of this runs (§7).

2 Lineage and contribution

Verity is downstream of, and indebted to, a specific body of work; naming it precisely is part of being honest about what is and is not new here.

What Verity builds on. The automatic comparison of bullet land impressions is the CSAFE / Iowa State method of Hare, Hofmann & Carriquiry [1]: striation signatures are extracted from each land, compared on *multiple* features (cross-correlation among them, alongside matching/non-matching striae counts and related measures), and combined by a trained random forest — that paper is explicit that cross-correlation alone does not suffice. The open `bulletxtrctr` pipeline and the three-similarity-score study of Vanderplas et al. [12] are the reference implementation and benchmark; Verity's Phase-1 score adopts the signature-extraction-and-correlation front end of that line. The cell-counting idea for impressed marks — partition one surface into cells, register each against the other, and count the cells whose best-fit translations and rotations agree — is Song's Congruent Matching Cells (CMC) [2] and the NIST error-rate line built on it [3], with `cmcR` as the open reference implementation [17]. The objective comparison of striated toolmarks by a distribution-free statistic is Chumbley et al. [4], improved and made deterministic by Hadler & Morris (the `toolmaRk` package) [18] and adapted to bullet lands by Krishnan & Hofmann [19]; open, fully automatic cartridge-case comparison is also due to Tai & Eddy [20]. The surface representation is the ISO 25178-72 X3P format, for which the reference reader is the `x3ptools` R package [11]. The evaluation metric C_{llr} and the calibration philosophy come from forensic speaker recognition, originally Brümmer & du Preez [13], with logistic-regression calibration as codified for forensic use by Morrison [21] and the validation regime for LR methods set out by Meuwly, Ramos & Haraksim [22].

What is novel in Verity. Five things, and we state them plainly so the boundary is clear:

1. **Unification.** One calibrated pipeline scores striated, impressed, and (in principle) fractured marks through *Congruent Matching Regions* (§3.4), a generalization of CMC from 2-D cells under translation+rotation to regions of any dimension under any transformation

group. CMC and the 1-D Chumbley statistic become two instantiations of one algorithm — and we *demonstrate* both reductions source-disjoint, under a single scorer configuration, alongside bullets (§5.2); the 3-D fractured case is the general form, not yet validated.

2. **A calibrated-LR answer to Cuellar et al.** Rather than a single disputed error rate, Verity reports a bounded LR with a per-study, source-disjoint C_{llr} (§5).
3. **A glass-box decision layer.** Region-level attribution (which parts of the surface drove the match) plus a *monotone, bounded* calibration — the "audit firewall" (§4) — so the reported evidence is interpretable regardless of how the underlying score was produced.
4. **Production engineering.** A native Rust X3P codec, Python and R bindings, a deployed calibrated-LR REST API, and a data catalog (§7).
5. **A small methodological result.** The `diag_contrast` scorer (§3.3), selected over the mean-diagonal baseline and over multivariate fusion by a barrel-disjoint ablation (§5.1).

Verity is the *deployment, calibration, and validation layer* for this research program — not a claim to have invented the comparison.

3 Method

The pipeline is four stages: surface metrology, region extraction, pairwise comparison, and aggregation. We describe the striated (bullet-land) path in full because it is the one we validate most heavily, then give the cross-modality generalization.

3.1 Surface metrology

A raw X3P scan is a height field $z(x, y)$ on a regular grid with invalid (masked) points. Individualizing texture lives in a narrow band of spatial wavelengths; everything coarser is the shape of the object and everything finer is measurement noise. Verity isolates the band with the standard ISO operations:

1. **Form removal** (the ISO 25178 F-operation). Fit a least-squares 2-D polynomial of total degree 2 to the valid points and subtract it, removing gross shape such as the curvature of a bullet land. With coordinates normalized to $[-1, 1]$ the design uses the monomials $x^i y^j,\ i + j \le 2$.
2. **Roughness-band isolation** (ISO 16610). An *S-filter* (Gaussian low-pass at λ_s) removes short-wavelength noise; an *L-filter* (subtract the λ_c Gaussian mean line) removes waviness. The residual is the height in the band $[\lambda_s, \lambda_c]$. The ISO 16610 Gaussian has 50% transmission at the cutoff, fixing

$$\sigma = \frac{\alpha\,\lambda}{\sqrt{2\pi}}, \qquad \alpha = \sqrt{\tfrac{\ln 2}{\pi}} \quad (\sigma \approx 0.1874\,\lambda), \tag{1}$$

and all filtering is NaN-aware (normalized convolution), so masked points neither contribute nor are filled. Deployed cutoffs are $\lambda_s = 4\,\mu\text{m}$ and $\lambda_c = 250\,\mu\text{m}$.

3.2 Region extraction: orientation and the signature

A bullet land carries signal only in its interior: the two groove shoulders at the ends of the across-striae axis have roughly 25× the amplitude of the striae and are common to every land, so they swamp a naive cross-correlation. Verity isolates the striae band:

1. **Orient.** Estimate the striae direction from the 2-D power spectrum. A Hann window suppresses the rectangular-edge cross; a mid-frequency annulus ($0.04\,r_{\max} < r < 0.5\,r_{\max}$) drops the DC term and the low-frequency grooves; the dominant power direction is read off the spectral second-moment tensor, and the striae run perpendicular to it. The field is rotated so the striae are vertical.

2. **Collapse.** Average along the striae (weighted by validity) to the 1-D across-striae profile.
3. **Crop.** Remove a fixed fraction off each end (default: keep the inner 50%) to discard both groove shoulders.

The cropped across-striae profile is the *signature* — the object compared. This FFT-based orientation is the single most consequential engineering choice in the pipeline: replacing a structure-tensor orientation with it took the barrel-disjoint C_{llr} on Hamby-252 from $\approx$0.60 to $\approx$0.11 and recovered the wide-scan studies (Phoenix, Hamby-173) from chance to AUC ≈ 0.96.

3.3 Pairwise comparison and the land×land matrix

Two signatures are compared by their peak normalized cross-correlation over integer lags. For mean-centred signals a, b,

$$\text{ncc}(a,b) = \max_{\ell} \frac{\sum_t a(t)\, b(t-\ell)}{\|a\|\,\|b\|} \in [-1,1]. \tag{2}$$

A bullet carries several lands, so comparing two bullets means comparing every land of one against every land of the other. This yields a land×land matrix CCF $\in [-1,1]^{n\times m}$ with $\text{CCF}_{ij} = \text{ncc}(a_i, b_j)$ and a matching integer-lag matrix. A genuine same-source pair lines up on a single cyclic diagonal: one land offset k makes *all* land pairs correlate at *consistent* lags. The mean correlation along cyclic offset k is

$$d(k) = \frac{1}{n}\sum_{i=1}^{n} \text{CCF}_{i,\,(i+k) \bmod m}, \qquad k^\star = \arg\max_k d(k), \tag{3}$$

and $d(k^\star)$ is the `diag_mean` score — the Phase-1 bullet score, identical to the `bulletxtrctr` "average of the diagonal" [12].

The Phase-1 score throws away exactly the structure that separates a real match from a lucky maximum-over-offsets. Verity computes the full matrix once and exposes that structure. Writing $\mathcal{D} = \{(i, (i+k^\star) \bmod m)\}$ for the winning diagonal, the deployed score is the *diagonal contrast*

$$\texttt{diag\_contrast} = \underbrace{\frac{1}{n}\sum_{(i,j)\in\mathcal{D}} \text{CCF}_{ij}}_{\text{matched diagonal}} - \underbrace{\frac{1}{|\overline{\mathcal{D}}|}\sum_{(i,j)\notin\mathcal{D}} \text{CCF}_{ij}}_{\text{background level}}, \tag{4}$$

i.e. how far the matched diagonal stands above the rest of the matrix. It is near zero for different-source pairs (no special diagonal) and large for same-source pairs. Verity also computes the runner-up offset margin $d(k^\star) - d(k^{\star\star})$, the lag coherence $1/(1+\text{std(diagonal lags)})$, and the weakest diagonal land — features that the ablation in §5.1 tests but which the deployed system does not require.

3.4 Congruent Matching Regions: one algorithm across modalities

Song's Congruent Matching Cells partitions one impressed surface into a grid of cells, registers each cell against the other surface over translation and rotation, and counts the cells whose best-fit transforms cluster [2, 3]. Verity observes that nothing in the counting argument is specific to 2-D cells or to the translation+rotation group. *Congruent Matching Regions* (CMR) lifts it to:

> *partition surface A into regions; register each region against B over a transformation group G; count the regions whose best-fit transforms agree — the CMR count.*

A genuine same-source pair has many regions that *independently* agree on one geometry; a non-match's regions register at scattered offsets, so no large congruent cluster forms. Because

Modality	Region	Transform group G	Specialist analogue
Striated (bullet land, toolmark)	1-D profile window	translation (lag)	Chumbley / consecutive striae [4]
Impressed (breech face, firing pin)	2-D cell	translation + rotation	Song CMC [2, 3]
Fractured / 3-D	3-D patch	rigid motion $SE(3)$	(general case)

Table 1. Congruent Matching Regions instantiated per modality. The 2-D row instantiates CMC-style cell counting; the 1-D row instantiates lag-agreement (consecutive-matching-striae-style) counting. One implementation, three marks; §5.2 measures how closely each generic reduction approaches its tuned specialist.

only *some* regions need to match, CMR is robust to partial, damaged, or warped marks where a single global cross-correlation is brittle. The core is domain-agnostic: it sees only (transform, correlation) votes and returns the largest cluster whose transforms agree within a tolerance. Each modality supplies a *vote producer* (Table 1).

The congruent regions are not only a score: they are the examiner-facing evidence of *which* parts of the surface drove the match, returned as overlay coordinates on both marks — the attribution half of the glass box.

4 From score to evidence: the calibration firewall

A score — whether the cross-correlation `diag_contrast` of today or a learned embedding similarity tomorrow — is not yet evidence. Verity turns a score into a likelihood ratio through a *monotone, fitted* calibration with published diagnostics. Because the map is a one-dimensional monotone transform, the reported LR is interpretable and auditable *no matter how the score was produced*: this is the firewall that keeps the decision out of any black box.

Score-based likelihood ratios, and their caveat. Verity reports a *score-based* LR: it calibrates a scalar similarity against the empirical same-source and different-source score distributions of a reference population, rather than modelling the full feature densities (a "feature-based" LR). This is the pragmatic, widely used choice, but the literature is clear that a score-based LR is a property of *the score and the chosen reference*, not an unconditional measure of the evidence, and that it can behave poorly if the reference is unrepresentative or the score discards relevant information [14, 15]. Verity's response is threefold: name the reference explicitly, bound the LR to what that reference can support (below), and report a credible interval that propagates the reference's finiteness.

Calibration map (Platt scaling). The default calibration is a two-parameter logistic fit (Platt scaling; logistic-regression calibration in the forensic-LR literature [21]) of the same-source posterior on the score. Two deliberate settings make it the right *deployment* calibration: the fit is near-unregularized (so the calibration slope is not shrunk away), and classes are balanced — fit as if the prior is 0.5 — so the posterior odds *are* the likelihood ratio and the (C_{llr}) objective's implied equal prior is respected. An isotonic (pool-adjacent-violators) fit is also available; it is the most flexible monotone calibration and is used to compute the discrimination floor $C_{\text{llr}}^{\text{min}}$, but it overfits when trained on few sources and is not deployed.

An empirical cap on the LR (the firewall's safety rail). A near-unregularized logistic fit on a *small* calibration set emits over-confident LRs: a hard held-out pair landing in the class overlap can receive a catastrophically wrong, large-magnitude LR that dominates C_{llr}. Inspired by the empirical lower-and-upper-bound (ELUB) principle of Vergeer et al. [16], Verity caps $|\log_{10}\text{LR}|$ at what the data can support. The deployed cap is set from the rarer class count,

$$|\log_{10}\text{LR}| \;\le\; \log_{10}\big(\max(n_{\text{minority}}, 10)\big), \tag{5}$$

the intuition being that one cannot assert evidence stronger than about $n_{\text{same}}{:}1$ from n_{same} same-source examples. We call this the *empirical cap*, not ELUB, because it deliberately simplifies the cited procedure in two ways that should be visible: Vergeer et al. derive *direction-specific* bounds from the validation LRs themselves (the upper bound governed chiefly by the different-source sample, the lower by the same-source sample), whereas Eq. (5) is a single symmetric cap on the class counts; and the floor of 10 means a reference with fewer than ten minority-class pairs is capped *less* conservatively than the principle implies. Implementing the direction-specific ELUB procedure on the validation LRs is planned. The cap is *monotone*, so the audit-firewall property (LR is a monotone transform of the score) is preserved, and it is tuning-free: it adapts to each reference's size automatically. Barrel-disjoint, the cap roughly halves the calibration loss (mean C_{llr} across the four bullet studies $0.44 \to 0.28$) while leaving the near-perfectly-discriminated Hamby-252 essentially untouched.

Diagnostics (C_{llr} and its decomposition, ECE). The system is summarized by the log-likelihood-ratio cost [13]

$$C_{\text{llr}} = \tfrac{1}{2}\left[\frac{1}{N_{\text{KM}}}\sum_{\text{same}}\log_2\!\left(1+\tfrac{1}{\text{LR}}\right) + \frac{1}{N_{\text{KNM}}}\sum_{\text{diff}}\log_2(1+\text{LR})\right], \tag{6}$$

where $C_{\text{llr}} = 0$ is perfect and $C_{\text{llr}} = 1$ is an uninformative system that always reports $\text{LR} = 1$. It decomposes as $C_{\text{llr}} = C_{\text{llr}}^{\min} + (C_{\text{llr}} - C_{\text{llr}}^{\min})$: $C_{\text{llr}}^{\min}$ (the cost after PAV-optimal calibration) is the *discrimination* floor — the best any monotone calibration of these scores could achieve — and the gap $C_{\text{llr}} - C_{\text{llr}}^{\min}$ is the *calibration* loss. Verity also reports the expected calibration error (ECE) of the LRs' implied posteriors, the equal error rate (EER), and Tippett plots. The verbal equivalents attached to reported LRs follow the ENFSI evaluative-reporting bands [23], and the validation regime — accuracy, discrimination, and calibration characterized on named data — follows the guideline of Meuwly, Ramos & Haraksim [22].

An interval on the LR. Because the score→LR map is itself estimated from a finite reference, a point LR hides its own uncertainty. Verity reports a *bootstrap percentile interval on* $\log_{10}\text{LR}$ — a resampling confidence interval, not a posterior credible interval (the API payload retains the historical field name `lr_credible_interval`) — by resampling the reference and *refitting the same bounded calibration in every replicate*, so the firewall holds per draw and the interval inherits the monotone, bounded property. Two schemes are provided: *row-stratified* (resample within each class, holding the class sizes and hence the empirical cap fixed, isolating calibration noise) and *clustered / block* (resample whole sources — the honest choice once references carry a source column, since pairs from one barrel are correlated and a row bootstrap understates the variance). The interval is a percentile interval because the LR distribution is asymmetric and piles at the cap; the reported point is the full-reference LR, not the bootstrap mean, to avoid resampling bias on the headline number.

The applicability-domain scope guard. Finally, the firewall has a gate. A calibration is valid only on data resembling its reference. Verity exposes the *calibrated domains* it is willing to score (currently: striated bullet lands, impressed breech-face marks, and striated screwdriver toolmarks) and declines to emit an LR for a scan outside them, rather than silently extrapolating a calibration past its evidence. This is the operational form of "refuse to claim a universal error rate."

5 Validation

The source-disjoint protocol. The central methodological failure Cuellar et al. [7] identify is that a system tested on the same sources it was tuned on, or evaluated with set-based designs

Quantity	Value	Notes
In-sample (full Hamby-252, deployed `diag_contrast`)		
AUC	0.999	rank separability
C_{llr}	0.077	calibration cost (Eq. 6)
$C_{\text{llr}}^{\text{min}}$	0.032	discrimination floor (PAV)
ECE	0.031	posterior reliability
EER	2.1 %	equal error rate
Empirical LR cap	$\lvert\log_{10}\text{LR}\rvert \approx 1.66$	Eq. 5
Barrel-disjoint test (full-study totals: 10 barrels, 46 KM / 549 KNM)		
Test C_{llr}	0.113 ± 0.066	mean ± SD over folds
Test AUC	1.000 ± 0.001	mean ± SD over folds

Table 2. Hamby-252 validation for the deployed `diag_contrast` scorer. The in-sample block characterizes the fitted system; the barrel-disjoint block is the honest generalization estimate. We report the latter as the headline. KM/KNM counts are full-study pair totals; each fold's test set comprises only the pairs among its four held-out barrels.

that leak class structure, reports an optimistic number that says nothing about a new firearm. Verity's primary protocol is therefore *barrel-disjoint* (source-disjoint) cross-validation: in each fold the calibration is fit on a subset of barrels and evaluated only on comparisons among the *held-out* barrels, so no barrel appears in both training and test. We report results *per study* and never pool across firearm makes, because pooling across populations is precisely the move that manufactures a misleading single error rate. (The deployed interactive reference does pool the four bullet studies for calibration breadth — but its different-source pairs are within-study only, it is labelled in-sample, and no validation claim rests on it; see §5.1.) Reported C_{llr} values are the mean and standard deviation across folds. Two properties of the fold design should be read alongside the SDs: the ten folds are repeated random source-level holdouts (40% of sources per fold, fixed seed) rather than a disjoint partition, so folds overlap in their held-out sources and the cross-fold SD is a stability measure, not an independent-replicate standard error; and each fold's test set is small in the source dimension (four held-out barrels on the ten-barrel sets; as few as 1–4 same-source pairs per fold on Fadul), so per-fold metrics carry wide sampling uncertainty of their own.

Datasets. Striated bullets are validated on the Hamby consecutively-rifled 10-barrel test sets (252 and 173) [8] and on two NBTRD [10] collections (a PGPD Beretta set and a Phoenix PD Ruger P-95 set). Impressed marks are validated on the Fadul study of 10 consecutively manufactured cartridge-case slides (breech faces) [9]. Striated *toolmarks* are validated on the MIT-licensed `tmaRks` consecutively-manufactured slotted-screwdriver set [24] (580 marks across 56 tool edges), a non-firearm domain the pipeline was never tuned for. Consecutively manufactured / rifled sources are the hardest available test: successive barrels or slides share the most subclass character, so a method that separates them is being held to the demanding case.

Headline result (Hamby-252, deployed `diag_contrast`). Table 2 reports the full Hamby-252 validation. The barrel-disjoint test $C_{\text{llr}} = 0.113 \pm 0.066$ (46 same-source / 549 different-source comparisons over 10 held-out barrels) with test AUC 1.000 ± 0.001 is the number we stand behind: it is measured on barrels the calibration never saw.

Per-study results. Table 3 and Figure 1 give the barrel-disjoint test C_{llr} across all four bullet studies, contrasting the deployed `diag_contrast` with the `diag_mean` baseline. The deployed scorer improves the calibration loss on three of the four studies; the NBTRD wide-scan sets are visibly harder than the Hamby benchmarks, exactly as one would expect, and the fold-to-fold

Study	Barrels	KM / KNM	Barrel-disjoint test C_{llr}	
			`diag_mean`	`diag_contrast`
Hamby-252 ([8])	10	46 / 549	0.121	**0.113**
PGPD Beretta (NBTRD [10])	10	30 / 405	0.299	**0.273**
Phoenix Ruger (NBTRD [10])	8	24 / 252	**0.252**	0.354
Hamby-173 ([8])	10	46 / 549	0.456	**0.338**

Table 3. Barrel-disjoint test C_{llr} per study (lower is better); **bold** marks the better scorer. The deployed `diag_contrast` wins on three of four; Phoenix Ruger is the exception. AUCs (`diag_contrast`): Hamby-252 1.000, PGPD Beretta 0.999, Phoenix Ruger 0.972, Hamby-173 0.971. Per-fold SDs are shown in Figure 1.

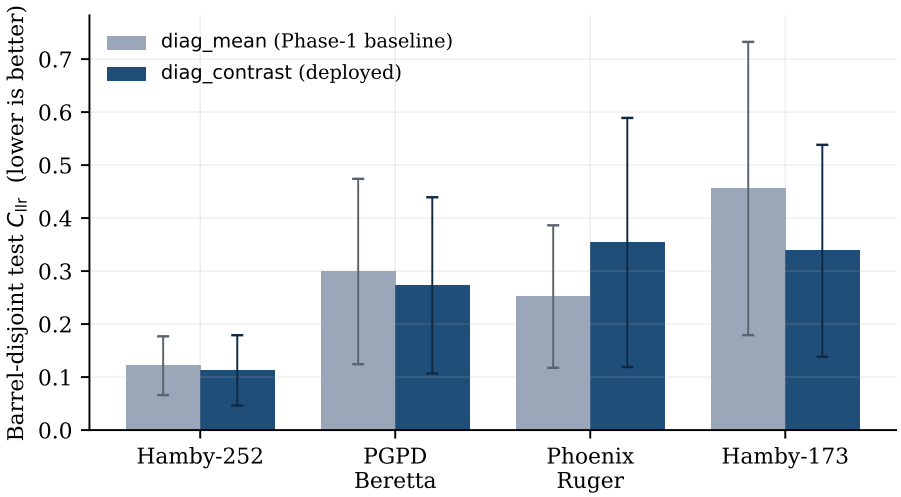

Figure 1. Barrel-disjoint test C_{llr} per study, baseline (`diag_mean`) vs deployed (`diag_contrast`); error bars are the cross-fold standard deviation. Generated from `data/ablation.json`. The wide error bars on the smaller NBTRD sets are the honest signature of limited source-disjoint data.

standard deviations are large on the smaller sets — a fact we surface rather than hide (§6).

Against the bullet specialist. The field-standard bullet method — the trained random-forest matchscore of Hare, Hofmann & Carriquiry [1], as implemented in `bulletxtrctr` [12] — run on the same scans through the identical barrel-disjoint fold protocol, reaches $C_{\text{llr}} = 0.064 \pm 0.015$ on Hamby-252 and 0.171 ± 0.088 on PGPD Beretta (AUC ≈ 1.000 on both; `data/baselines.json`). The trained specialist remains ahead of Verity's untrained `diag_contrast` (0.113 and 0.273 respectively), as one would expect — Verity's contribution on bullets is the calibrated, bounded, deployable LR layer, not a better matcher — and the same head-to-head is published on the project site.

5.1 Scorer selection by ablation

The choice of `diag_contrast` as the production scorer was made *by the data*, not assumed. We ran a barrel-disjoint ablation over the candidate scalars derived from the CCF matrix — the `diag_mean` baseline, `diag_contrast`, the offset margin, lag coherence, the weakest diagonal land — and two multivariate fusions of those features (a standardized logistic, and an isotonic fusion). The result is a genuine rigor highlight, and it cuts against the reflex to add model capacity:

- `diag_contrast` beats the multivariate logistic fusion on three of the four studies, and has the lowest *mean* barrel-disjoint C_{llr} across studies (Figure 2, Table 4);

Scorer	Hamby-252	PGPD	Phoenix	Hamby-173
`diag_mean` (Phase-1)	0.121	0.299	0.252	0.456
`diag_contrast` (deployed)	0.113	0.273	0.354	0.338
fusion (logistic)	0.127	0.298	0.231	0.456
fusion (isotonic)	0.110	0.294	0.273	0.531

Table 4. Barrel-disjoint test C_{llr} by scorer and study (from `data/ablation.json`). `diag_contrast` has the lowest cross-study mean (Fig. 2); the isotonic fusion's strong fit does not survive the source-disjoint split.

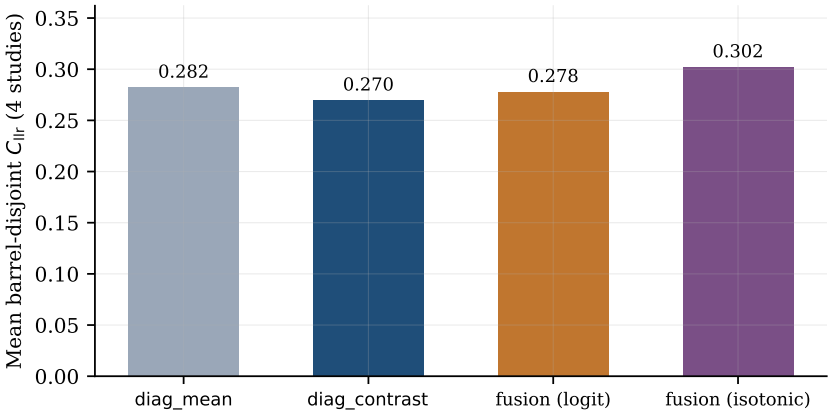

Figure 2. Mean barrel-disjoint C_{llr} across the four bullet studies for each candidate scorer. `diag_contrast` (deployed) is lowest; adding fusion capacity does not help on source-disjoint data, and the isotonic fusion is worst.

- the isotonic fusion — the most flexible option — *overfits*: it has the worst mean source-disjoint C_{llr}, despite the strongest in-sample discrimination;
- so the simple, single-number, fully interpretable `diag_contrast` is the deployed scorer. A multivariate fusion remains available as an opt-in research path with two or three inspectable coefficients — but it is not the default, because the data did not justify it.

The pooled deployed reference. For interactive use the deployed API ships a pooled striated reference of $n_{\text{KM}} = 146$ same-source and $n_{\text{KNM}} = 1755$ different-source land comparisons; on it the fitted system reaches AUC ≈ 0.984, $C_{\text{llr}} \approx 0.193$, and $C_{\text{llr}}^{\text{min}} \approx 0.168$ (pooling the four studies, including the harder NBTRD sets, raises this above the per-study figures). This is an *in-sample* characterization of the bundled calibration and is labelled as such; it is not a generalization claim. The generalization claims are the barrel-disjoint numbers above.

5.2 Closing the CMR table: cross-modality validation

Table 1 is a *claim*: one algorithm, three reductions. Here we *demonstrate* it. The same congruent-matching pipeline, under the *same* scorer configuration (one content hash shared across all three calibration references), is validated source-disjoint on each modality (Table 5). Every figure is recomputed from a committed reference by the identical source-disjoint path used for bullets; the comparison baselines are measured by a re-lockable harness (§7).

Impressed (Fadul, CMR-2D). The deployed 2-D reduction — the same `cmr_count` consensus, now over 2-D cells under translation+rotation — reaches in-sample AUC 0.997 (C_{llr} 0.220, $C_{\text{llr}}^{\text{min}}$ 0.070) and a slide-disjoint C_{llr} of 0.385 ± 0.187 on the ten consecutively manufactured Fadul

Reduction (modality)	Reference	Src-disjoint C_{llr}	AUC	Baseline / specialist
Bullet lands (1-D striae)	pooled, 4 studies	0.186 ± 0.126	0.989	bulletxtrctr 0.064; Hamby-252 alone 0.113
Cartridge breech (2-D)	Fadul, 10 slides	0.385 ± 0.187	0.991	cmcR 0.194; areal 0.529
Screwdriver toolmark (1-D)	`tmaRks`, 56 edges	0.328 ± 0.050	0.957	global align 0.447

Table 5. The CMR table, *demonstrated*: one scorer configuration (shared config hash) reduced to each modality and validated source-disjoint (C_{llr} lower is better; mean ± SD over folds). Bullets are the pooled four-study reference (Hamby-252 alone is 0.113); cartridges and toolmarks are their full sets. CMR matches or approaches the field-standard specialist in each case, and on toolmarks *beats* the same-pipeline global baseline. From `data/cmr_table.json` and `data/baselines.json`.

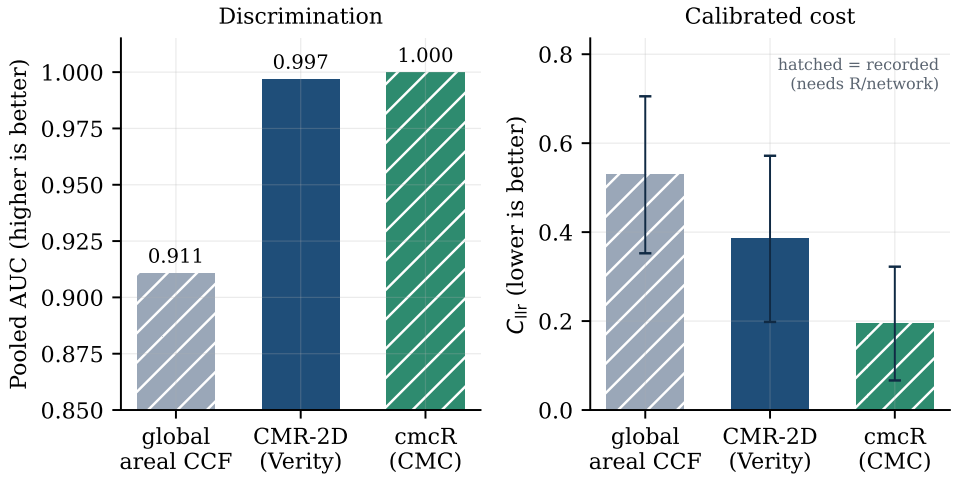

Figure 3. Impressed-mark validation on the slide-disjoint Fadul task: the naive global areal CCF, the deployed generic CMR-2D, and the `cmcR` CMC specialist. Left: pooled AUC. Right: source-disjoint C_{llr} with cross-fold error bars. CMR-2D (solid; reproducible from the committed reference) lands beside the specialist; the baselines (hatched) need R to re-lock. Generated from `data/cartridge_fadul.json` and `data/baselines.json`.

slides. Figure 3 places it against the two references on the *same* slide-disjoint task: generic CMR-2D moves the naive global areal cross-correlation (pooled AUC 0.911, C_{llr} 0.529) most of the way to the `cmcR` Congruent Matching Cells specialist [17] (AUC 1.000, C_{llr} 0.194) — with no cartridge-specific engineering. CMR-2D instantiates the same counting principle as CMC; the remaining gap to the tuned specialist (C_{llr} 0.385 vs 0.194) is the measured price of keeping the implementation generic, and closing it is ongoing work, not a claim already made.

Toolmarks (`tmaRks`, CMR-1D). The 1-D reduction — the consecutive-striae statistic of Table 1 — is validated on the MIT `tmaRks` consecutively-manufactured screwdriver set [24] (580 marks across 56 tool edges), source-disjoint by the mark-generating edge. (A flat-head screwdriver carries two working edges; the split is disjoint by *edge* — the mark generator — so same-tool cross-edge pairs are treated as different-source. The harness also supports tool-level-disjoint splits; locking that sensitivity analysis is planned.) CMR-1D reaches AUC 0.957 at C_{llr} 0.328 ± 0.050 and, notably, *beats* the same-pipeline global 1-D cross-correlation on the identical set (AUC 0.942, C_{llr} 0.447): requiring congruence among matching windows, rather than one global alignment, helps on real toolmarks. This is a *non-firearm* domain the pipeline was never tuned for — evidence the unification is not firearm-specific. (On the much smaller Ames Lab set, our run of the `toolmaRk` U statistic [18] gives pooled AUC 0.614 over 7 tools; with so few tools and replicates, that set is a weak benchmark for *any* method, which is why we report `tmaRks` as the primary toolmark benchmark.)

Reading these numbers. The cross-modality C_{llr} values are larger than Hamby-252's 0.113, as they should be: cartridges and toolmarks have far fewer sources (ten slides; the cross-edge toolmark task is genuinely hard), and the pooled bullet reference mixes the harder NBTRD sets. The claim is not that every modality is as strong as the best bullet study; it is that *one calibrated algorithm, validated source-disjoint, matches or approaches the field-standard method in each of the three reductions of Table 1* — Chumbley-style counting on toolmarks, CMC-style counting on breech faces, and the bullet-land result — rather than three bespoke pipelines.

6 Limitations and honest scope

We would rather state the boundaries plainly than have a reviewer find them.

- **Model selection used the validation studies; no untouched confirmation set exists yet.** The deployed scorer (`diag_contrast`), the FFT orientation front-end (§3.2), and the filter cutoffs were chosen by barrel-disjoint ablations on the *same* four bullet studies reported in §5. Each fold is honest — the calibration never sees a held-out source — but repeated evaluation against the same source-disjoint splits during development means the headline numbers are best read as *development-set performance under a source-disjoint protocol*, not as a one-shot confirmation on new data. This is the same selection-leakage failure mode Cuellar et al. [7] flag in black-box studies, and it applies to us until a confirmation set is run. The frozen open-benchmark splits commit the protocol going forward; a one-shot validation on data never used in development (new NBTRD collections; the LAPD corpus) is the next validation milestone.
- **Small source-disjoint sets widen uncertainty.** Consecutively manufactured benchmarks are small in the dimension that matters — sources. Fadul has ≈ 10 slides; the bullet sets have 8–10 barrels. With so few held-out sources the fold-to-fold C_{llr} has a large standard deviation (Fig. 1), and the empirical LR cap (Eq. 5) is correspondingly low. More independent sources would tighten both.
- **Score-based LR limitations.** As in §4, the reported LR is a property of the score and the named reference, not an unconditional weight of evidence [14, 15]; a reference unrepresentative of a casework population would mis-calibrate, which is why the scope guard exists.
- **Not a universal error rate.** Nothing here should be read as "the error rate of firearms identification." Every number is conditioned on a named population and a stated protocol. This is the intended reading of our answer to Cuellar et al., not a hedge.
- **The learned representation does not yet beat the baseline.** A Phase-2b learned (embedding) representation is implemented behind the same firewall, but on the data available it does not yet outperform the cross-correlation baseline. We treat this as a data limitation, not a success, and the deployed system uses the interpretable baseline. The concrete path is self-supervised pretraining on the unlabelled LAPD ($\approx$15k-scan) corpus, fine-tuned behind the same firewall and retested barrel-disjoint — an orthogonal research track, not a precondition for the calibrated-LR results above.
- **The cross-modality reductions rest on few sources.** The impressed (Fadul, ≈ 10 slides) and toolmark (`tmaRks`) reductions are now demonstrated source-disjoint (§5.2) — and CMR-1D even beats the same-pipeline global baseline on toolmarks — but with far fewer independent sources than the bullet benchmarks, so the wider C_{llr} and low LR caps reflect that. They are a genuine cross-modality demonstration of the CMR unification, not yet a definitive per-modality error characterization.

7 The platform

The methods above are only useful in court and in the lab if they are reproducible, fast, and inspectable. Verity is built as production software, not a script.

- **Native Rust X3P codec.** A from-scratch reader/writer for the ISO 25178-72 / ISO 5436-2 X3P format (`crates/verity-x3p`), with Python and R bindings. It interoperates with the reference `x3ptools` R reader [11] but removes the heavy runtime dependency, giving fast, portable scan I/O for both research and deployment.
- **A calibrated-LR REST API.** A deployed service exposes `/detect` (suggest a mark type from a scan), `/compare` (return the calibrated, bounded LR with its bootstrap interval and the region attribution), and a `/health` endpoint that publishes the currently calibrated domains — the scope guard, machine-readable.
- **A local-first data catalog.** A catalog service harvests and labels the open NBTRD [10] 3-D scans — which have no public API — succeeding the earlier CSAFE scraper (unmaintained since 2020) with a maintained, reproducible harvester, alongside the Fadul and `tmaRks` sets, so validation runs against versioned, content-addressed, reproducible data.
- **An open, frozen benchmark.** Frozen source-disjoint splits (bullets, cartridge, toolmark) with published split hashes, a downloadable replication kit whose offline scorer matches the server exactly, and a public leaderboard ranked by total C_{llr}. The leaderboard protocol scores one LR per pair via a single leave-the-pair's-sources-out calibration, which differs from the per-fold refits reported above — e.g. Fadul source-disjoint C_{llr} is 0.385 under §5.2's protocol and 0.398 on the leaderboard. Both are source-disjoint; the kit reproduces the leaderboard numbers exactly.
- **Reproducible validation reports.** Every number in §5 is recomputed from a committed calibration reference by a checked-in harness and persisted as the paper's source of truth: `data/ablation.json` (bullets), `data/cmr_table.json` (the three reductions), `data/cartridge_fadul.json`, `data/toolmark_tmaRks.json`, and `data/baselines.json` (the comparison baselines, re-lockable offline via `verity-relock-baselines`); the figures are regenerated from those files by a checked-in script.

An invitation. Verity is the production, calibration, and validation layer for a research program it did not invent. The natural next step is to validate it on larger, more representative, genuinely source-disjoint reference collections — the kind of data the CSAFE program is built to produce. We offer the codec, the calibration firewall, and the validation harness as open infrastructure, and we would welcome a collaboration that puts a properly powered reference behind the calibrated LR. That is a proposal, not a claim of endorsement.

Acknowledgments

Verity stands entirely on the CSAFE / Iowa State research program and the NIST congruent-matching line: the bullet-matching method of Hare, Hofmann & Carriquiry [1], the `bulletxtrctr` pipeline and similarity-score study of Vanderplas et al. [12], the Congruent Matching Cells method of Song and colleagues [2, 3], the Chumbley toolmark statistic [4], the `x3ptools` format tooling [11], the open benchmark data [8, 9, 10], and the calibration metric of Brümmer & du Preez [13]. Mention of CSAFE denotes intellectual lineage and a prospective collaboration only; it does not imply review or endorsement of Verity by CSAFE, Iowa State University, NIST, or any author of the cited works.

References

[1] E. Hare, H. Hofmann & A. Carriquiry (2017). Automatic matching of bullet land impressions. *The Annals of Applied Statistics* **11**(4), 2332–2356. doi:10.1214/17-AOAS1080.

[2] J. Song (2013). Proposed congruent matching cells (CMC) method for ballistic identification and error rate estimation. *AFTE Journal* **45**(2), 184–194. NIST.

[3] J. Song, T. V. Vorburger, W. Chu, J. Yen, J. A. Soons, D. B. Ott & N. F. Zhang (2018). Estimating error rates for firearm evidence identifications in forensic science. *Forensic Science International* **284**, 15–32. doi:10.1016/j.forsciint.2017.12.013.

[4] L. S. Chumbley, M. D. Morris, M. J. Kreiser, C. Fisher, J. Craft, L. J. Genalo, S. Davis, D. Faden & J. Kidd (2010). Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm. *Journal of Forensic Sciences* **55**(4), 953–961. doi:10.1111/j.1556-4029.2010.01424.x.

[5] National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press. doi:10.17226/12589.

[6] President's Council of Advisors on Science and Technology (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Executive Office of the President.

[7] M. Cuellar, S. Vanderplas, A. Luby & M. Rosenblum (2024). Methodological problems in every black-box study of forensic firearm comparisons. *Law, Probability and Risk* **23**(1), mgae015. doi:10.1093/lpr/mgae015.

[8] J. E. Hamby, D. J. Brundage & J. W. Thorpe (2009). The identification of bullets fired from 10 consecutively rifled 9 mm Ruger pistol barrels: a research project involving 507 participants from 20 countries. *AFTE Journal* **41**(2), 99–110.

[9] T. G. Fadul, G. A. Hernandez, S. Stoiloff & S. Gulati (2011). An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides. *AFTE Journal* **43**(3). NIJ Award 2009-DN-BX-K230.

[10] X. Zheng, J. Soons, T. V. Vorburger et al. (NIST) (2020). NIST Ballistics Toolmark Research Database. *Journal of Research of NIST* **125**, 125004. doi:10.6028/jres.125.004.

[11] H. Hofmann, S. Vanderplas, G. Krishnan & E. Hare (2024). x3ptools: Tools for Working with 3D Surface Measurements. R package (CRAN), v0.0.4. doi:10.32614/CRAN.package.x3ptools.

[12] S. Vanderplas, M. Nally, T. Klep, C. Cadevall & H. Hofmann (2020). Comparison of three similarity scores for bullet LEA matching. *Forensic Science International* **308**, 110167. doi:10.1016/j.forsciint.2020.110167.

[13] N. Brümmer & J. du Preez (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language* **20**(2–3), 230–275. doi:10.1016/j.csl.2005.08.001.

[14] A. B. Hepler, C. P. Saunders, L. J. Davis & J. Buscaglia (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International* **219**(1–3), 129–140. doi:10.1016/j.forsciint.2011.12.009.

[15] C. Neumann & M. Ausdemore (2020). Defence against the modern arts: the curse of statistics — Part II: "Score-based likelihood ratios". *Law, Probability and Risk* **19**(1), 21–42. doi:10.1093/lpr/mgaa006.

[16] P. Vergeer, A. van Es, A. de Jongh, I. Alberink & R. Stoel (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: when to stop extrapolating? *Science & Justice* **56**(6), 482–491. doi:10.1016/j.scijus.2016.06.003. (Empirical-bound / ELUB principle.)

[17] J. Zemmels, S. Vanderplas & H. Hofmann (2023). A study in reproducibility: the congruent matching cells algorithm and `cmcR` package. *The R Journal* **14**(4), 79–102. RJ-2023-014.

[18] J. R. Hadler & M. D. Morris (2018). An improved version of a tool mark comparison algorithm. *Journal of Forensic Sciences* **63**(3), 849–855. (Implemented in the `toolmaRk` R package.)

[19] G. Krishnan & H. Hofmann (2019). Adapting the Chumbley score to match striae on land engraved areas (LEAs) of bullets. *Journal of Forensic Sciences* **64**(3), 728–740. doi:10.1111/1556-4029.13950.

[20] X. H. Tai & W. F. Eddy (2018). A fully automatic method for comparing cartridge case images. *Journal of Forensic Sciences* **63**(2), 440–448. doi:10.1111/1556-4029.13577.

[21] G. S. Morrison (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences* **45**(2), 173–197. doi:10.1080/00450618.2012.733025.

[22] D. Meuwly, D. Ramos & R. Haraksim (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International* **276**, 142–153. doi:10.1016/j.forsciint.2016.03.048.

[23] S. M. Willis et al. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science* (Strengthening the Evaluation of Forensic Results across Europe, v3.0). European Network of Forensic Science Institutes. enfsi.eu.

[24] M. Cuellar, S. Gao & H. Hofmann (2024). An algorithm for forensic toolmark comparisons. *Forensic Science International: Synergy*. arXiv:2312.00032. Data: the `tmaRks` consecutively-manufactured screwdriver toolmark scans, github.com/heike/tmaRks (MIT license).