Forensic Chemistry 49 (2026) 100746

Contents lists available at ScienceDirect

Forensic Chemistry

journal homepage: www.sciencedirect.com/journal/forensic-chemistry

DyeSPY-LINK: The first likelihood-based inference of near-source kinship for dyed hair evidence comparisons

Aidan P. Holman [a,b], Dmitry Kurouski [a,b,*,1]

[a] Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, United States
[b] Interdisciplinary Faculty of Toxicology, Texas A&M University, College Station, TX, United States

ARTICLE INFO

Keywords:
Forensic analysis
Likelihood
LR
SLR
Hair dye
Hair

ABSTRACT

Hair and fiber evidence is among the most frequently encountered and environmentally persistent forms of trace material in forensic casework. Within this category, forensic practitioners consistently recognize that treated hair, such as bleached or dyed, provides substantially greater probative value than untreated hair. At the same time, modern forensic practice is moving toward likelihood ratio (LR)-based interpretation to ensure transparent, quantitative assessments of evidential strength. Despite this shift, no LR-based framework has yet been established for dyed hair. Here we introduce DyeSPY-LINK, a score-based LR (SLR) system for the probabilistic evaluation of near-source kinship between dyed hairs. Specifically, DyeSPY-LINK is designed to determine the strength of evidence for whether two dyed hair samples originate from the same dye. Performance was assessed on 17 nonoxidative and 43 oxidative commercial dyes previously characterized in DyeSPY. Using surface-enhanced Raman (SER) spectra, we construct known-match (KM) and known-non-match (KNM) comparisons between dyed-hair sources and transform cosine similarity scores into SLRs via kernel density estimation (KDE). For nonoxidative dyes, KM and KNM score distributions were well separated (AUC = 0.9990) coupled with a very low log-likelihood ratio cost (CLLR, 0.0325) indicating excellent discrimination and near-perfect calibration. Misleading-evidence rates were below ~1% across a wide range of thresholds, and each additional shared colorant increased the odds of obtaining strong support for a common source (LR ≥ 100) by approximately 2.2-fold. For oxidative dyes, KM and KNM distributions were also well separated (AUC = 0.9616) with higher, but still appropriate CLLR (0.3298). Misleading evidence rates remained low, and each additional shared colorant increased the odds of strong support for a common source by 2.6-fold. DyeSPY-LINK thus lays the groundwork for a fully probabilistic, empirically validated system for interpreting dyed-hair evidence in forensic casework.

1. Introduction

The interpretation of forensic evidence has undergone a substantial shift in the past two decades, moving away from categorical conclusions and toward probabilistic frameworks that quantify the strength of evidence [1]. Central to this evolution is the likelihood ratio (LR), which expresses how much more strongly the observed data support one proposition over another. LR-based approaches have been widely adopted in areas such as forensic DNA analysis, fingerprints, handwriting, voice comparison, and increasingly in chemical and trace evidence examinations [2]. Their appeal stems from the LR's ability to incorporate population variability, quantify evidential strength on a continuous scale, and transparently separate the roles of the scientist (evaluating evidence) and the court (evaluating propositions) [3]. As forensic disciplines advance toward greater statistical rigor and standardization, LR-based interpretation frameworks are becoming essential tools for ensuring defensible, transparent, and reproducible expert evaluations.

At the same time, dyed hair represents an important but historically underutilized form of trace evidence. More than half of the global market for dyestuff (including lipsticks, blushers, eye shadow, and nail polish) is hair dyes [4]. Coincidentally, dyed hairs are frequently encountered in violent crimes, assaults, and interpersonal transfer scenarios [5]. Unlike untreated hair, dyed hair contains exogenous colorant molecules that absorb into or bind to the keratin fiber, leaving chemically rich signatures that persist over time [6]. These dye molecules vary

* Corresponding author at: Department of Biochemistry and Biophysics, Texas A&M University College Station, TX, United States.
E-mail address: dkurouski@tamu.edu (D. Kurouski).
[1] ORCID Dmitry Kurouski: 0000–0002–6040-4213.

https://doi.org/10.1016/j.forc.2026.100746
Received 29 December 2025; Received in revised form 2 April 2026; Accepted 9 April 2026
Available online 10 April 2026

across commercial formulations, manufacturers, and oxidative versus nonoxidative dyeing pathways. As a result, dyed hairs carry a complex chemical profile that can serve as a discriminating feature in forensic investigations; provided that appropriate analytical and interpretive tools are available [7,8]. Traditional methods such as UV–Visible spectrophotometry, Infrared and Raman spectroscopy, and mass spectrometry have shown potential but often lack the sensitivity or high-throughput interpretive framework needed to translate spectral differences into objective, case-relevant evidence assessments [9].

The DyeSPY tool was developed to address this gap by combining vibrational spectroscopy with machine learning to identify (i) hair-dye type (oxidative vs. nonoxidative), (ii) the constituent colorants present in a dyed fiber, and (iii) the perceptual color category of the treated hair [7]. These three phases provide a nondestructive, high-throughput platform for characterizing dyed hair using surface-enhanced Raman spectroscopy (SERS) coupled with statistical learning models. However, while DyeSPY enables accurate classification and characterization of hair colorants, its current framework does not quantify the strength of evidence when comparing two dyed hairs. Specifically, whether they are more likely to originate from the same or different colorant mixtures is currently unexplored. Such a capability is crucial in forensic contexts where investigators or courts may ask whether two questioned hairs arise from a common source or exhibit chemical similarity beyond that expected by chance [3].

The present study introduces DyeSPY-LINK, an LR-based probabilistic module designed to extend DyeSPY from a classification platform to a forensic evidence interpretation system. DyeSPY-LINK uses spectral similarity metrics derived from SER spectra, conditioned on permanence reactivity (such as in Phase I DyeSPY), to model the expected variation within and between colorant mixtures. By estimating the probability density of similarity scores for same-mixture (or known-match, KM) and different-mixture (or known-non-match, KNM) comparisons using kernel density estimation (KDE), the system computes score-based LRs (SLRs) that quantify the strength of evidence for or against a shared mixture origin; in other words, whether two dyed hairs originate from the same dye. Furthermore, we are not questioning whether two dyed hairs originate from the same individual since hair dye, the key component of SER spectra of dyed hair, is circumstantial and not individualizing evidence. This analytic LR approach avoids the statistical pitfalls associated with classifier-based SLR and ensemble-SLR methods and provides a transparent, chemically interpretable evidential quantity aligned with modern forensic LR guidelines [10].

In doing so, this work establishes the first LR-based statistical interpretation system for dyed hair evidence. It demonstrates how spectral chemistry, machine learning colorant predictions, and probabilistic evaluation can be combined to provide robust, objective, and interpretable evidence relevant to source inference. The resulting framework enables forensic practitioners to articulate not only what colorants are present in a dyed hair, but also how strongly two hairs support a hypothesis of common or different dye mixture origins; advancing the role of dyed hair as a scientifically grounded and impactful form of trace evidence.

2. Materials and methods

2.1. Bayesian framework for forensic source inference

Generally, hair evidence is considered probative [11]. Forensic practitioners are encouraged to use probabilistic estimations when interpreting probative evidence [3]. This is done using the Bayes Theorem, Fig. 1. Briefly, the responsibility of the forensic practitioner is to evaluate the evidence and generate a likelihood (the LR) for observing two items of evidence originating from the same source (H_s, prosecutor's hypothesis) or originating from different sources (H_d, defense's hypothesis) [3,12]. Thus, we will build a framework for rigorously evaluating the LR that two samples share the same colorant compositions.

For practical purposes, likelihood ratios can be estimated through score-based LRs (SLRs), which reduce complex spectral data into a univariate (single) similarity score [10]. In this framework, an LR is approximated as the ratio of the densities of these scores under the same-mixture and different-mixture conditions:

$$LR \approx SLR = \frac{P(data|same\ mixture)}{P(data|different\ mixture)} = \frac{P(score|same\ mixture)}{P(score|different\ mixture)}$$

Scores can be generated using regression models (e.g., partial least squares, random forests) or by parametric or nonparametric similarity measures [13–15]. In this study, we adopt a nonparametric, similarity-based approach, using cosine similarity to compute initial similarity scores. This allows us to make minimal assumptions about the behavior of our data.

Unlike full classifier-based LR systems, similarity scores alone do not incorporate upstream classification uncertainty (e.g., Phase I oxidation decisions, Phase II mixture predictions, or estimation of the source). A common strategy to reflect such uncertainty is random resampling (e.g., Monte Carlo, bootstrap) of KM and KNM pairs [13].

Here, KM comparisons were constructed exhaustively by pairing spectra originating from the same dye source, ensuring complete characterization of within-source variability. KNM comparisons were generated by randomly pairing spectra from different dye sources using label shuffling, with sampling performed to maintain approximate balance between KM and KNM sets. These empirical similarity scores are then modeled using KDE to obtain continuous estimates of $P(\delta|KM)$ and $P(\delta|KNM)$, from which score-based likelihood ratios are derived (a visual summary of the SLR computational workflow is demonstrated in Scheme 1):

$$SLR(\delta) = \frac{f_{KM}(\delta)}{f_{KNM}(\delta)} = \frac{P(\delta|KM)}{P(\delta|KNM)}$$

Furthermore, for (S)LRs to be valid, two fundamental assumptions must be held. First, the propositions must be mutually exclusive: any given comparison must belong to either the KM population or the KNM population, but never both. Second, the propositions must be collectively exhaustive: KM and KNM must cover all logically possible origins for the observed pair, such that no relevant scenario exists outside these two defined populations. These conditions are naturally satisfied in DyeSPY-LINK, where mixture identity is explicitly defined and all KM and KNM comparisons are unique, enumerated or resampled accordingly.

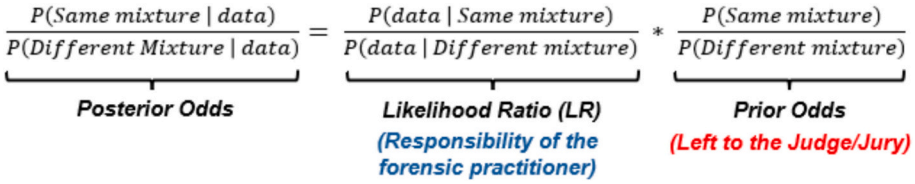
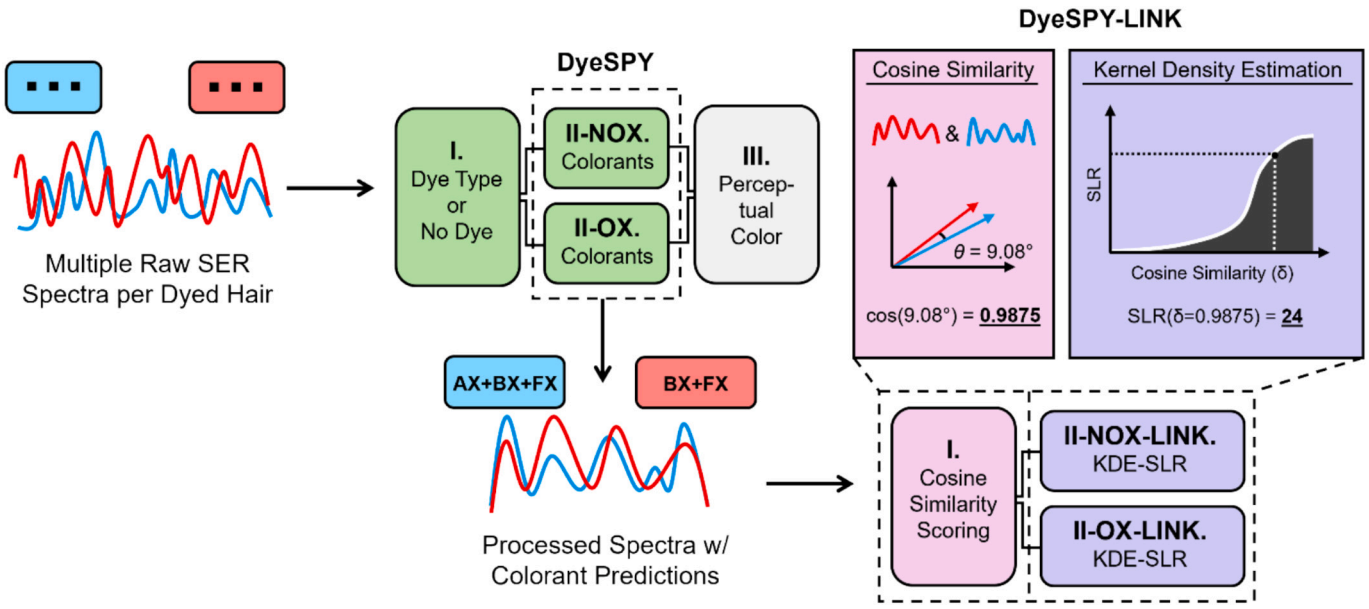


Fig. 1. Bayes theorem and responsibilities in a forensic context.



Scheme 1. SLR computational workflow of DyeSPY-LINK.

2.2. Sample selection

To evaluate the performance of the DyeSPY-LINK module for both nonoxidative and oxidative dyes, all corresponding dyed hair spectra from the original DyeSPY database were analyzed [7]. These datasets can be accessed via a free public repository here: https://zenodo.org/records/18881839.

In the initial study, five spectra were collected from each of three hair strands per dye, meaning each dye represented three independent sources. Thus, the 17 nonoxidative dyes correspond to 51 total sources (5 spectra/items per source), and the 43 oxidative dyes correspond to 129 sources (5 spectra per source). Notably, prior likelihood-based evaluations for forensic science have often used (far) fewer sources and items than those available here [13,16,17].

2.3. Data analysis

All spectra were trimmed to the 450–1650 cm^{-1} range to reduce edge noise, baseline-corrected using the asymmetric least squares algorithm ($\lambda = 1E+5$, $p = 0.01$), smoothed using a first-order Savitzky-Golay filter (window length = 7, polyorder = 1), and area-normalized prior to analysis, as done in the original DyeSPY study [7]. This was done using Python 3.13 and the following libraries: NumPy, pandas, SciPy, pybaselines, and scikit-learn.

Cosine similarity was computed in Python as the normalized dot product between two processed spectra. To model the empirical score distributions under KM and KNM conditions, we applied univariate Gaussian kernel density estimation (KDE) using scipy.stats.gaussian_kde. Bandwidths, h, were selected automatically using Scott's rule ($h = \sigma n^{-1/5}$), yielding KM and KNM-dependent kernel widths for each model. Scott's rule for bandwidths provides a data-driven and asymptotically optimal smoothing parameter for unimodal distributions, determined objectively without manual tuning. Stability of KDE estimates was assessed via resampling, confirming minimal variation in density shape and resulting SLR distributions across subsets of the data.

Regarding performance-linked measurements, the log-likelihood ratio cost (CLLR) and the receiver operating characteristic (ROC) area under the curve (AUC) were used to assess the robustness of the LR estimations. CLLR was computed using custom Python routines following Brümmer and du Preez's formulation [18], while AUC was obtained using roc_auc_score from scikit-learn. CLLR quantifies LR calibration, that is, how well the magnitudes of the LRs reflect the underlying KM and KNM score distributions. The minimum CLLR ($CLLR_{min}$) measures the best discrimination performance when ignoring calibration error. The AUC measures discriminative ability by estimating the probability that a randomly chosen KM comparison receives a higher LR than a randomly chosen KNM comparison.

For statistical significance, a non-parametric one-sided sign test (OSST; $\alpha = 0.05$) was used to evaluate whether KM and KNM SLRs were significantly above and below 1.0, respectively.

Logistic regression was used to model how a factor (e.g., number of shared colorants) influences the probability of a defined outcome, such as obtaining strong evidential support (e.g., LR $\geq$ 100), by relating it to the log-odds of that outcome. In practice, the model estimates how each unit increase in the predictor changes the odds of strong support, allowing practitioners to interpret results as multiplicative changes in likelihood (e.g., "each additional shared colorant increases the odds of strong support for same-source hypothesis by ~2-fold"). The regression coefficients, representing the log-odds of a predictor given the outcome, were assessed for significant difference from zero using the Wald test ($\alpha = 0.05$), with a non-significant result indicating that the predictor does not reliably influence the odds.

3. Results and discussion

3.1. Nonoxidative dyes

Nonoxidative dye colorants are primarily composed of direct dyes, which typically behave in a linearly additive fashion in SER spectra [7]. As a consequence, spectra from dyes that share one or more colorants tend to exhibit systematic similarity, whereas spectra from mixtures with disjoint colorant sets diverge more substantially. This structural property implies that a continuous measure of spectral similarity should differentiate KM from KNM pairs, albeit with some overlap due to shared chemistry.

To quantify this relationship, we used cosine similarity as the initial scoring function. Cosine similarity measures the angle between two high-dimensional vectors, rather than their absolute magnitudes, and is widely used in domains where relative pattern shape matters more than absolute intensity; such as text embedding models [19], speaker recognition embeddings [20,21], and mass spectrometry fingerprinting [22,23]. By focusing on spectral shape rather than raw amplitude, cosine

similarity provides a stable and normalized similarity score: high values (close to 1.0) indicate closely aligned spectral signatures whereas lower values (close to 0) indicate dissimilar mixtures (of dyes) or dye compositions. Unlike distance-based metrics, it is relatively insensitive to magnitude differences among corresponding peaks, making it well suited for dyed hair evidence where environmental exposure (e.g., sunlight) can attenuate chromophore signal without substantially altering spectral structure [8]. This makes it an appropriate and interpretable first-stage quantifier of source proximity for both KM and KNM comparisons.

Building on these similarity scores, we next applied KDE to model the empirical distribution of KM and KNM scores. KDE is a nonparametric density estimation technique that constructs a smooth approximation of a probability distribution by placing a small kernel (e.g., Gaussian) at each observed data point. KDE can be thought of as drawing a smooth curve over observed similarity scores to show where values are most common and filling in the gaps for scores with no existing pairs using neighboring scores with existing pairs. In this study, one curve represents similarities expected for same-dye comparisons (KMs, expressed as $f_{KM}[\delta]$), and another represents different-dye comparisons (KNMs, $f_{KNM}[\delta]$). The SLR for a given cosine similarity score is then calculated as the ratio of the KM density to the KNM density at that score. KDE has been applied extensively in score-based likelihood ratio systems, including those for forensic voice comparison [24], fingerprint similarity [25], handwriting analysis [14,16], and more recently chemical profile matching [26,27]. Its flexibility allows the model to capture complex, multimodal score distributions without assuming a particular parametric family, unlike logistic regression-based modeling.

For example, consider two dyed hair spectra with a cosine similarity of 0.9120. To evaluate the strength of evidence, we examine how frequently this level of similarity occurs among KMs compared to KNMs. If the estimated densities at this score are 60 for KMs and 2 for KNMs, the resulting SLR is 30, indicating that this degree of similarity is 30 times more likely under the same-dye proposition than under the different-dye proposition, given the reference dataset. In practice, this would generally be interpreted as support for a common dye source instead of odds-based due to model-data specificity, while acknowledging that the strength of this support should be considered alongside performance characteristics such as misleading-evidence rates. Score-based LR systems of this form are standard in modern forensic evaluation frameworks and align with best practices recommended by the European Network of Forensic Science Institutes (ENFSI), the Aitken-Taroni Bayesian paradigm, and applications in chemometric mixture attribution and spectroscopic evidence interpretation [3,10,15,28].

A total of 2400 known-match (KM) comparisons were generated based on the full set of nonoxidative dye sources, and, to maintain approximate balance while still reflecting realistic variability in non-match space, the number of known-non-match (KNM) comparisons was fixed at 2000. Using cosine similarity as the scalar scoring function, KM pairs exhibited substantially higher similarity values than KNM pairs, as expected (Table 1). Additionally, KNMs exhibited substantially greater variability (relative standard deviation; RSD = 31.17%) compared to KMs (RSD = 1.54%). This observation is consistent with vibrational spectroscopy principles, wherein different chromophores possessing similar functional groups can produce partially overlapping spectral features, leading to elevated similarity scores relative to chemically dissimilar species. These similarity scores were then modeled using KDE to obtain empirical score distributions which were subsequently converted into LRs via the standard SLR mapping.

Table 1
Summary statistics of cosine similarity scores for KM and KNM comparisons for nonoxidative colorant mixtures.

Group	Mean	Standard deviation	Sample size
KMs	0.9898	0.0153	2400
KNMs	0.5406	0.1685	2000

The resulting LR system demonstrated excellent discrimination between KM and KNM comparisons. The AUC was 0.9990, indicating that the system assigns higher SLRs to KM pairs than to KNM pairs in more than 99.9% of all possible KM-KNM pairings. Calibration performance was similarly strong: the empirical CLLR was 0.0325, well within the range considered "very good" for calibrated forensic LR systems and consistent with or better than KDE-based SLR estimation reported in other score-based domains (e.g., speaker recognition and trace-evidence profiling) [2,29]. Additionally, the $CLLR_{min}$ was also 0.0325, indicating that our SLR system is already at its theoretical optimum.

The LR distribution summaries further reinforce this behavior. KM SLRs ranged from approximately 2.97 to 765.21, with a median substantially above 1 (median = 582.97, $p < <0.001$, OSST), indicating systematic support for the KM hypothesis when dyes share constituent colorants, Fig. 2. In contrast, KNM LRs spanned approximately <0.001 to 168.27, with the majority concentrated well below 1 (median < 0.001, $p < <0.001$, OSST), providing expected support for the non-match hypothesis when dyes differ in colorant composition (Fig. S1 displays KDE distribution curves). Misleading-evidence analysis showed low rates of LR misassignment across threshold levels, with $m_{KM}\left(SLR \leq \frac{1}{T}\right)$ below 0.1% and $m_{KNM}(SLR \geq T)$ consistently small, ~1%, and monotonic for $T \in \{2, 3, 5, 10, 20\}$, Table 2. These trends indicate both very low false-exclusion and low false-inclusion tendencies, which are central to the defensibility of LR-based conclusions in forensic comparison settings [3].

One may ask what the effect of the number of shared colorants is on the magnitude of SLRs. We would expect that as the number of shared colorants, n_{shared}, increases, the SLR increases by a factor, β (The relative frequency of KM and KNM comparisons for each n_{shared} can be found in Fig. S2). To determine this, we used logistic regression to understand how shared colorants influence evidential strength. The formula can be written as $Log_{10}(SLR_i) = \alpha + \beta^* n_{shared,i} + \varepsilon_i$ where α represents the baseline or intercept and ε_i represents the model error for a given comparison, i. To determine the odds of an effect, E, given the n_{shared}, we rewrite our equation as $Logit(P(E)) = \alpha + \beta^* n_{shared}$, which gives us the log-odds, that can then be converted into a linear odds ratio using e^{β}. To determine our E, we consulted previous literature on what LR thresholds are acceptable. Both the National Institute of Standards and Technology (NIST) and ENFSI recommend that evidential thresholds be selected in regions where $m_{KM}\left(SLR \leq \frac{1}{T}\right)$ and $m_{KNM}(SLR \geq T)$ are generally small, leaving the final choice to the practitioner. In our data, misleading-evidence rates remained low across all evaluated thresholds, and LRs ≥100 are commonly interpreted as providing strong or moderately strong support for H_s [15,30]. Accordingly, we adopted SLR ≥ 100 as the evidential threshold, E.

The final fit model is $Logit(P(SLR \geq 100)) = -1.048 + 0.795^* n_{shared}$. The β of 0.795 (SE = 0.032, Wald $p < <0.001$) can be converted to an odds ratio per +1 shared colorant of 2.21. So, each additional shared colorant increased the odds of obtaining SLRs ≥ 100 by ~2.2 fold. Correspondingly, the probability of SLR ≥ 100 rose from 0.26 (0 shared colorants) to 0.89 (up to 4 shared colorants), starting with 2 shared colorants possessing a probability greater than chance (0.632), Table 3. When we flip our support for the defense, i.e. SLR ≤ 1/100, our model becomes $Logit(P(SLR \leq 1/100)) = 2.916 - 4.129^* n_{shared}$. Our new odds ratio of 0.02 ($\beta = -4.129$, SE = 0.123, Wald $p < <0.001$) tells us that each additional shared colorant lowers the likelihood of obtaining SLRs ≤ 100 by a factor of 50-fold. Consequently, the probability of SLR ≤ 1/100 dropped from 0.95 to 0.23 starting at 1 shared colorant. This demonstrates that colorant-level overlap strongly influences the magnitude of similarity-based likelihood ratios, supporting our hypothesis.

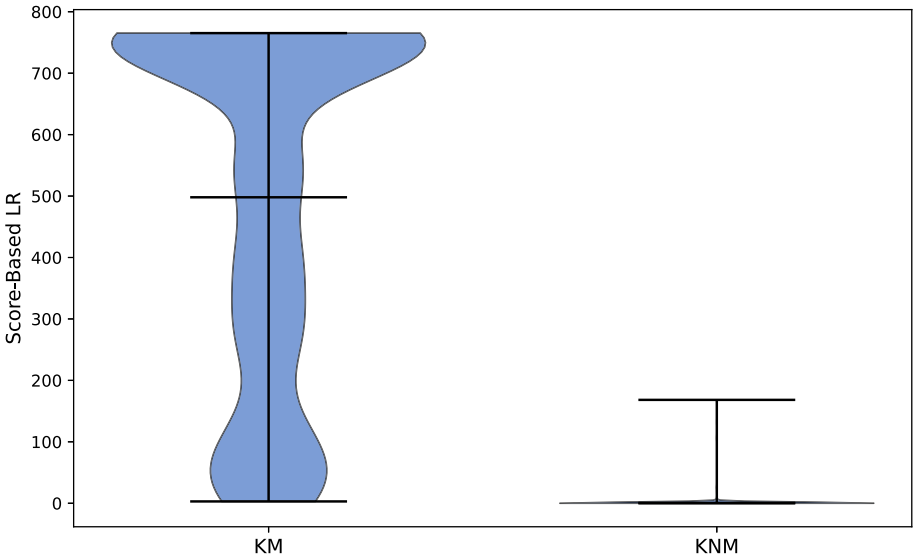


Fig. 2. Score-based LR distributions for KM and KNM nonoxidative colorant mixture comparisons. Horizontal bars show mean and IQR.

Table 2
Misleading evidence rates by threshold (T) in nonoxidative colorant mixture SLR estimations.

T	$m_{KM}\left(SLR \leq \frac{1}{T}\right)$	$m_{KNM}(SLR \geq T)$
2	<0.001	0.009
3	<0.001	0.009
5	<0.001	0.009
10	<0.001	0.007
20	<0.001	0.006
100	<0.001	0.004
150	<0.001	0.001
200	<0.001	<0.001

Table 3
Logistic regression predicted probabilities for obtaining SLR ≥ 100 and ≤ 1/100 by the number of shared colorants in nonoxidative colorant mixtures.

n_{shared}	$P(SLR \geq 100)$	$P(SLR \leq 1/100)$
0	0.260	0.949
1	0.437	0.229
2	0.632	0.005
3	0.792	<0.001
4	0.894	<0.001

The SLR system for nonoxidative dyes showed exceptionally strong performance across KM and KNM comparisons. Cosine similarity and KDE modeling produced clear separation between match and non-match distributions, yielding near-perfect discrimination and excellent calibration. KM pairs consistently generated large SLRs, KNM pairs produced LRs well below 1, and misleading-evidence rates remained below 1% across thresholds. Logistic modeling further showed that each additional shared colorant more than doubled the odds of strong support for a match. Overall, the system provides highly reliable evidential weighting for nonoxidative colorant mixtures.

3.2. Oxidative dyes

As is well established in the dye chemistry literature (and as motivated in our original rationale for treating oxidative and nonoxidative systems separately) the behavior of oxidative colorants in mixture form is fundamentally distinct from that of direct (nonoxidative) dyes [7]. Oxidative dyes consist primarily of primary intermediates and couplers which, in the presence of an activating agent such as hydrogen peroxide or ammonia, undergo a series of redox-driven coupling reactions that generate entirely new chromophores. These reaction products possess altered electronic environments and markedly different patterns of polarizability, yielding Raman signatures that differ substantially from those of their precursor molecules.

This chemical transformation has direct implications for mixture interpretation. Because oxidative dyes generate unique reaction products whose compositions depend on the specific combination of intermediates, couplers, and oxidizer-to-species ratios, cosine similarity estimates for oxidative dyed hairs are theoretically more stable and source-informative. For instance, if dye Alpha contains primary intermediate *p*-phenylenediamine (PPD), whereas dye Beta contains primary intermediate PPD and coupler resorcinol (RES), oxidation will yield high-abundance products such as Bandrowski's base (BB) in Alpha, a trimer of PPD, and predominantly 3-Hydroxy-*N*-(4-aminophenyl)-*p*-benzoquinone monoimine, or red indoaniline dye (IAD), a dimer of PPD and RES in Beta, with further polymerization contingent on oxidizer-to-species ratio. [31] These distinct reaction pathways produce correspondingly different SER spectra (Fig. 3).

A total of 3900 KM and 4000 KNM comparisons were generated based on the full set of oxidative dye sources. Using cosine similarity as the scalar scoring function, KM pairs exhibited higher similarity values than KNM pairs, as expected (Table 4). These similarity scores were subsequently modeled using KDE to obtain empirical score distributions which were subsequently converted into LRs via the standard SLR mapping.

The resulting SLR system demonstrated excellent discrimination between KM and KNM comparisons. The AUC was 0.9616, indicating that the system assigns higher SLRs to KM pairs than to KNM pairs in more than 96% of all possible KM-KNM pairings. The empirical CLLR was 0.3298, falling squarely within the range generally interpreted as "good" calibration for forensic likelihood-ratio systems, and comparable to the performance reported for KDE-based SLR estimation in other fields [13]. Importantly, the $CLLR_{min}$ was 0.3292, indicating that our SLR system is effectively identical to its theoretical optimum. Notably, the cost is higher than in the nonoxidative SLR estimator. We attribute this to the greater number of possible shared colorants, up to 7, Fig. S3.

The KDE-generated SLR distributions for KMs ranged from 0.073 to 252.75, with a median substantially above 1 (median = 205.22, p < <0.001, OSST). The SLR distribution for KNMs was from <0.001 to 194.96, with a median below 1 (median = 0.180, p < <0.001, OSST), Fig. 4 (Fig. S4 displays KDE distribution curves). Misleading-evidence

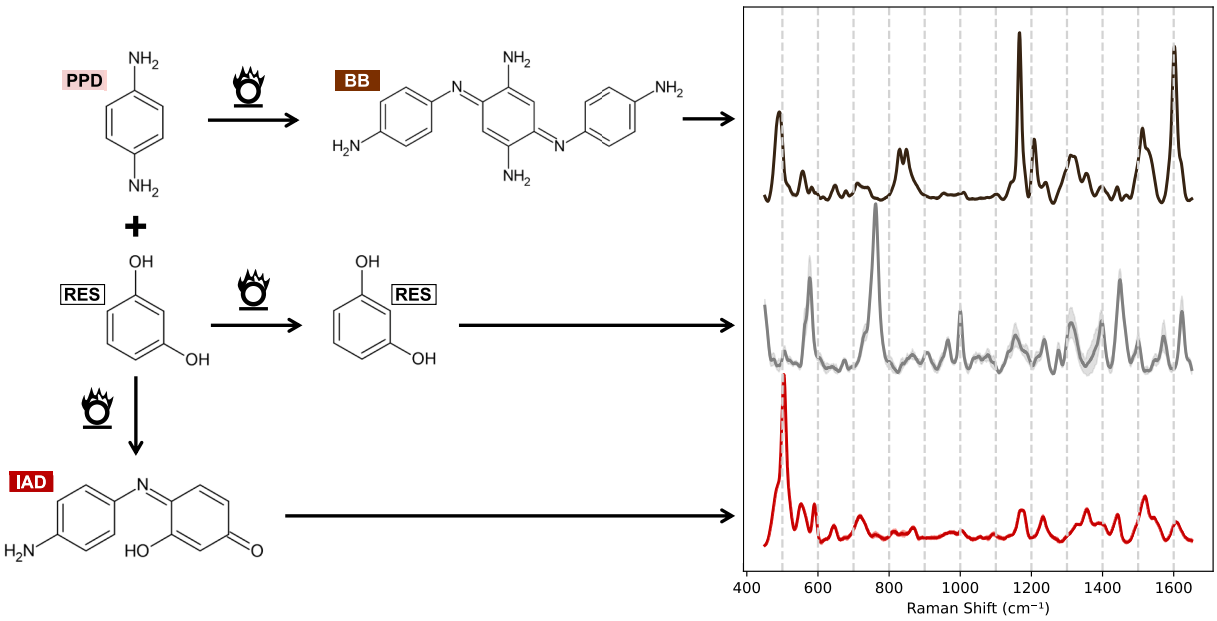


Fig. 3. Oxidative dye chemistry is species and concentration-dependent. PPD oxidized alone vs. in the presence of a coupler yields different products dominating the SER signal. SER spectra represent mean and SD of 2% (*w/v*) PPD, RES, and PPD with RES (top to bottom) mixed with hydrogen peroxide (2%, final concentration) and left to react for 15 min. Small changes, including lateral shifts among peaks, can be observed in SER spectra of BB and IAD.

Table 4
Summary statistics of cosine similarity scores for KM and KNM comparisons for nonoxidative colorant mixtures.

Group	Mean	Standard deviation	Sample size
KMs	0.9548	0.0818	3900
KNMs	0.7148	0.1075	4000

analysis showed low rates of LR misassignment across threshold levels, with $m_{KM}\left(SLR \leq \frac{1}{T}\right)$ below 10% starting at SLR of 3 and $m_{KNM}(SLR \geq T)$ consistently small, ~1%, and monotonic for $T \in \{2, 3, 5, 10, 20\}$, Table 5. These trends indicate both low false-exclusion and false-inclusion tendencies.

Given that, for both oxidative and nonoxidative mixture SLR estimations, a decision threshold of 10 consistently yields misleading-evidence rates below 1%, and a threshold of 100 yields misleading-evidence rates below 0.5%, we adopt the interpretive scale presented in Table 6. These thresholds and their corresponding verbal categories are aligned with internationally accepted guidance from NIST, ENFSI, and the Association of Forensic Science Providers (AFSP) [3,30]. By way of illustration, an SLR of 20 may be communicated as follows: "In my opinion, the correspondence between the dyed hair recovered from the crime scene and the dyed hair taken from the accused provides support for the proposition that both hairs originated from the same dye or set of colorants, rather than from different dye sources." This phrasing is preferred over statements such as "the hairs are 20-times more likely to come from the same mixture than not," because such numerical interpretations imply assumptions that are not part of standard LR reasoning. In addition, similarity-based LR approaches, such as SLRs, yield values that are relative and model-dependent, and thus should not be presented as direct probabilistic claims about the underlying

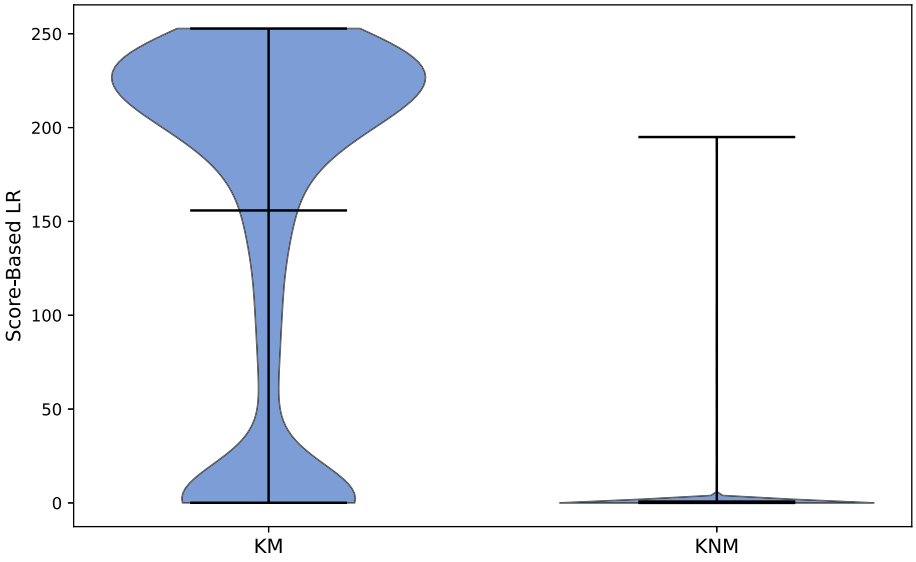


Fig. 4. Score-based LR distributions for KM and KNM oxidative colorant mixture comparisons. Horizontal bars show mean and IQR.

Table 5
Misleading evidence rates by threshold (T) in oxidative colorant mixture SLR estimations.

T	$m_{KM}\left(SLR \leq \frac{1}{T}\right)$	$m_{KNM}(SLR \geq T)$
2	0.115	0.014
3	0.068	0.010
5	0.014	0.008
10	0.001	0.007
20	<0.001	0.005
100	<0.001	0.001
150	<0.001	0.001
200	<0.001	<0.001

Table 6
Suggested forensic SLR interpretation.

Benchmark	Interpretation
$SLR \leq 0.01$	Strong support for H_d
$0.01 < SLR \leq 0.1$	Support for H_d
$0.1 < SLR < 10$	Inconclusive
$10 \leq SLR < 100$	Support for H_s
$SLR \geq 100$	Strong support for H_s

propositions. Instead, the role of the expert is to explain the strength of the evidence in a measured, transparent way, consistent with established forensic reporting standards.

Using logistic regression, we modeled the relationship between the increase in shared colorants and the SLR probability. Both propositional logistic regression models can be found in Table S1. Unlike our non-oxidative dye library which possesses up to 4 shared colorants across dyes, our oxidative dye library contains mixtures with up to 7 shared colorants. We found that each additional shared colorant increased the odds of obtaining SLRs $\geq$ 100 by 2.6-fold. Correspondingly, the probability of SLR $\geq$ 100 rose from 0.02 (0 shared colorants) to 0.93 (up to 7 shared colorants), starting with 5 shared colorants possessing a probability greater than chance (0.67), Table 7. On the other hand, each additional shared colorant lowers the likelihood of obtaining SLRs $\leq$ 100 by a factor of 2.6-fold. Consequently, the probability of SLR $\leq$ 1/100 dropped from 0.56 to 0.34 starting at 1 shared colorant, further supporting that colorant-level overlap strongly influences the magnitude of similarity-based likelihood ratios, supporting our hypothesis.

The SLR system for oxidative dyes showed strong performance across KM and KNM comparisons. Cosine similarity and KDE modeling produced clear separation between match and non-match distributions, yielding an acceptable calibration. KM pairs consistently generated large SLRs, KNM pairs produced LRs well below 1, and misleading-evidence rates remained below 1% across thresholds starting at SLR = 10. Logistic modeling further showed that each additional shared colorant more than doubled the odds of strong support for a match. Overall, the system provides reliable evidential weighting for oxidative colorant mixtures.

Table 7
Logistic regression predicted probabilities for obtaining SLR $\geq$ 100 and $\leq$ 1/100 by the number of shared colorants in nonoxidative colorant mixtures.

n_{shared}	$P(SLR \geq 100)$	$P(SLR \leq 1/100)$
0	0.017	0.565
1	0.044	0.336
2	0.106	0.165
3	0.235	0.072
4	0.443	0.029
5	0.674	0.012
6	0.843	0.005
7	0.933	0.002

3.3. Modularization of DyeSPY-LINK

DyeSPY-LINK was designed to answer a straightforward but challenging forensic question: how similar two dyed-hair samples are, and how strongly does that similarity support a common source? Importantly, this framework is intended to evaluate shared dye origin rather than individual attribution, as SERS selectively probes dye chemistry without capturing individualizing characteristics, a distinction that is critical for appropriate interpretation in casework and cannot be overstated. To make usage possible outside the development environment, the method was modularized so it can be reused with new cases and new samples.

The key idea behind modularization is separating training from application. During training, DyeSPY-LINK learns how similar spectra behave when they truly originate from the same colorant mixture (KM comparisons) versus when they come from different mixtures (KNM comparisons). This behavior is summarized when KDE's map the SLRs following cosine similarity scoring. Once trained, the KDEs are serialized ("pickled") and stored as fixed model objects using the Pickle library in Python. These saved models can then be loaded and applied to any future evidence comparison involving colorant mixtures represented in the training data, without the need to recompute or retrain the densities. This greatly reduces computational cost and ensures consistent, reproducible SLR estimates across analyses.

To use DyeSPY-LINK on new evidence, the user only needs two class columns: (1) ID, a unique identifier for each sample (e.g., "Alpha", "Bravo", "Charlie") and (2) Labels, the list of predicted colorants for that sample, separated by a "+" if more than one are present (e.g., AX+CX). These columns and the corresponding processed spectral data are *now automatically saved after Phase II classification* using the DyedHairModule in the DyeSPY platform, so no file modification is necessary from the user.

Once this file is prepared, DyeSPY-LINK computes pairwise similarity between all samples and converts those similarities into SLRs using the trained KDE models, NOX-LINK and OX-LINK. An example output for DyeSPY-LINK, using NOX-LINK specifically, is shown in Fig. 5. The entire process requires no statistical tuning at the point of use; making it accessible to practitioners, even outside a research laboratory.

Modularizing DyeSPY-LINK transforms a complex forensic likelihood-ratio framework into a reusable inference tool. By separating KDE training (NOX-LINK and OX-LINK estimators) from downstream casework, the system ensures reproducibility and transparency while simplifying external application. Phase II classifier outputs are saved in a

```
DyeSPY-LINK Summary Report by Pair
==============================
…
Comparison: C vs. E
    Query colorants (C)          : HX
    Known colorants (E)          : HX
    Shared colorants             : 1

    Cosine similarity (δ)        : 0.9909
    SLR                          : 303
    Log10(SLR)                   : 2.482

    Interpretation               : Strong support for Hs (same mixture).
    Practitioner note            :
        ▪ Treat 'Inconclusive' SLRs (0.1 < SLR < 10) as having limited
          evidential value.
        ▪ Very small SLRs (≤ 0.01) strongly favor Hd; very large SLRs (≥
          100) strongly favor Hs.
        ▪ Always interpret SLRs alongside case context and DyeSPY Phase
          I/II outputs.
…
```

Fig. 5. Example output for DyeSPY-LINK using the NOX-LINK model. Sample C represents the questioned or query sample with predicted colorant HX and sample E, a known source sample either from a suspect or experiment, with colorant HX.

standardized format which allows DyeSPY-LINK to automatically determine shared colorants, compute cosine similarity scores, and calculate SLRs using the pre-trained KDE models. This modular design enables DyeSPY-LINK to be used flexibly across new samples, new dyes, and new forensic scenarios without retraining or recalibration.

4. Novelty

DyeSPY-LINK represents a substantive conceptual and practical departure from the original DyeSPY framework rather than an incremental extension of it. Whereas DyeSPY was designed to classify dyed hair by pathway, colorant composition, and perceptual color, DyeSPY-LINK introduces an entirely separate inference pipeline that, for the first time, enables explicit quantification of the strength of dyed hair evidence using SLRs. This distinction is critical: DyeSPY-LINK reframes dyed hair analysis from a classification problem into an evidential evaluation problem, aligning it directly with modern forensic interpretation principles.

Although the present implementation leverages the colorant labels produced by DyeSPY Phase II for convenience and consistency, none of the DyeSPY phases are technically required to generate SLRs. In practice, a user could preprocess spectra using the procedures described here and directly submit the resulting data to the OX-LINK and NOX-LINK models to compute SLRs under competing dye-type assumptions, thereby evaluating evidential strength without any prior classification step. This flexibility allows DyeSPY-LINK to function as a standalone LR engine or to be integrated into broader analytical workflows. Conversely, when DyeSPY Phase I—or any comparable dye-type classifier—is employed, its output can be used to guide model selection, enabling the analyst to condition the LR calculation on the most probable dye pathway. Together, these features establish DyeSPY-LINK as the first platform to provide a modular, likelihood-based framework for dyed hair evidence, capable of operating independently or synergistically with classification models while maintaining a clear separation between feature extraction, decision support, and evidential interpretation.

5. Limitations

Although DyeSPY-LINK represents the first likelihood-based framework for near-source kinship inference in dyed hair evidence, several limitations of the present study must be acknowledged.

First, the SLR system was developed and evaluated exclusively using spectra from the original DyeSPY reference database. These data were generated under controlled laboratory conditions on a single SERS platform, using a fixed acquisition protocol, a limited set of commercial dyes, and standardized sample preparation. As such, the performance metrics reported here reflect this specific experimental domain. They may not fully generalize to casework conditions in which laboratories employ different instruments, laser wavelengths, substrates, or preprocessing pipelines, or in which hairs have undergone environmental exposure, cosmetic treatments, or ageing processes that were not represented in the development dataset. External validation with truly independent dyed-hair samples, acquired in different laboratories and under realistic forensic conditions, is therefore essential before routine operational use.

Additionally, the current validation focuses on evidential strength at the level of mixture similarity, not at the level of individual people. DyeSPY-LINK assesses whether two dyed hairs are more likely to share a common colorant mixture than to arise from different mixtures, under the assumptions and population represented by the DyeSPY database. It does not estimate the frequency of that mixture in any target population, nor does it address questions about the probability that a particular person is the source of the hair. As with other LR systems, the interpretation of the LR in the broader case context depends on population representativeness, case-specific propositions, and information external to the model. Users must therefore avoid overinterpreting mixture-based LRs as person-level source attributions.

The current scope of DyeSPY-LINK should be viewed as an intentional pilot-scale implementation rather than a methodological shortcoming. At present, the system recognizes 60 dyes which is sufficient to demonstrate feasibility, repeatability, and likelihood-based interpretation while clearly defining the boundaries of inference expected in real casework. Such staged development is well established in forensic science. A clear precedent is the Combined DNA Index System (CODIS), developed by the Federal Bureau of Investigation, which began in the 1990s with a limited number of STR loci [32], small population datasets, and few participating laboratories, yet evolved through systematic expansion, population studies, and standardization into a globally relied-upon forensic platform [33]. This limitation is particularly relevant because different commercial dyes may share similar or identical colorants, often at varying concentrations due to formulation constraints, which may reduce discrimination in real-world forensic comparisons. In the same way, DyeSPY-LINK's restricted dye library functions as a foundational reference set for controlled validation and LR calibration, with anticipated expansion to broader commercial formulations and hair origins occurring through iterative, transparent growth rather than being a prerequisite for scientific validity.

Finally, this study does not yet provide a comprehensive Daubert-oriented characterization of error rates for the full DyeSPY pipeline. While CLLR, CLLRmin, AUC, and misleading-evidence rates offer a robust statistical description of LR behavior under the study conditions, a complete operational validation would additionally require end-to-end estimates of false-positive and false-negative rates for realistic propositions and decision thresholds, evaluated on *independent* data. Until such external validation and error-rate estimation are completed, DyeSPY-LINK should be viewed as a proof-of-concept and developmental framework that demonstrates the feasibility and promise of LR-based interpretation for dyed hair evidence, rather than a fully mature casework tool.

6. Conclusions

This work situates dyed hair within the broader shift in forensic science toward probabilistic, LR-based interpretation. Although dyed hairs are chemically rich and commonly encountered in casework, they have lacked a transparent, quantitative framework for evaluating source similarity. By combining SERS measurements, machine-learning permanence predictions from the related DyeSPY platform, and score-based likelihood ratios, DyeSPY-LINK provides such a framework, allowing practitioners to express the strength of evidence for shared dye-mixture origins in a manner consistent with modern Bayesian and international forensic guidelines. This work establishes a foundation for future extensions involving broader dye populations, environmental robustness, and full operational error-rate characterization. Specifically, future validation of DyeSPY-LINK should include evidence of robustness to environmental degradation and contamination, as well as sample variability, including different underlying hair samples per dye and expanding the current library of commercial dyes, to name a few examples. Nevertheless, DyeSPY-LINK demonstrates that dyed hair, which has been long underutilized despite its ubiquity, can be integrated into a defensible, quantitative LR framework and potentially contribute meaningfully to contemporary forensic practice.

CRediT authorship contribution statement

Aidan P. Holman: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Dmitry Kurouski:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project was supported by Award No. 2020-90663-TX-DU, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.forc.2026.100746.

Data availability

Data will be made available on request.

References

[1] A.M. Christensen, C.M. Crowder, S.D. Ousley, M.M. Houck, Error and its meaning in forensic science, J. Forensic Sci. 59 (1) (2014) 123–126, https://doi.org/10.1111/1556-4029.12275.
[2] S. van Lierop, D. Ramos, M. Sjerps, R. Ypma, An overview of log likelihood ratio cost in forensic science–where is it used and what values can we expect? Forensic Sci. Int. Synerg. 8 (2024) 100466 https://doi.org/10.1016/j.fsisyn.2024.100466.
[3] C. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, 2004.
[4] A. Tkaczyk, K. Mitrowska, A. Posyniak, Synthetic organic dyes as contaminants of the aquatic environment and their implications for ecosystems: a review, Sci. Total Environ. 717 (2020) 137222, https://doi.org/10.1016/j.scitotenv.2020.137222.
[5] Hair WSP, Washington State Patrol Forensic Laboratory Services Bureau, 2026.
[6] S.A. Da França, M.F. Dario, V.B. Esteves, A.R. Baby, M.V.R. Velasco, Types of hair dye and their mechanisms of action, Cosmetics 2 (2) (2015) 110–126, https://doi.org/10.3390/cosmetics2020110.
[7] A.P. Holman, A. Maalouf, D. Kurouski, DyeSPY: establishing the first forensic SERS reference for hair dye colorant evidence, Anal. Chem. (2025), https://doi.org/10.1021/acs.analchem.5c05023.
[8] A. Holman, D. Kurouski, The effects of sun exposure on colorant identification of permanently and semi-permanently dyed hair, Sci. Rep. 13 (1) (2023) 2168, https://doi.org/10.1038/s41598-023-29221-8.
[9] A.P. Holman, D. Kurouski, Surface-enhanced raman spectroscopy in forensic analysis, Rev. Anal. Chem. 43 (1) (2024) 20230079, https://doi.org/10.1515/revac-2023-0079.
[10] N. Garton, D. Ommen, J. Niemi, A. Carriquiry, Score-based likelihood ratios to evaluate forensic pattern evidence, arXiv preprint arXiv:2002.09470, 2020, https://doi.org/10.48550/arXiv.2002.09470.
[11] M. Airlie, J. Robertson, E. Brooks, Forensic hair analysis–worldwide survey results, Forensic Sci. Int. 327 (2021) 110966, https://doi.org/10.1016/j.forsciint.2021.110966.
[12] A. Caliebe, S. Walsh, F. Liu, M. Kayser, M. Krawczak, Likelihood ratio and posterior odds in forensic genetics: two sides of the same coin, Forensic Sci. Int. Genet. 28 (2017) 203–210, https://doi.org/10.1016/j.fsigen.2017.03.004.
[13] F. Veneri, D.M. Ommen, Ensemble learning for score likelihood ratios under the common source problem, Stat. Anal. Data Min. The ASA Data Sci. J. 16 (6) (2023) 528–546, https://doi.org/10.1002/sam.11637.
[14] M.Q. Johnson, D.M. Ommen, Handwriting identification using random forests and score-based likelihood ratios, Stat. Anal. Data Min. The ASA Data Sci. J. 15 (3) (2022) 357–375, https://doi.org/10.1002/sam.11566.
[15] A. Martyna, G. Zadora, T. Neocleous, A. Michalska, N. Dean, Hybrid approach combining chemometrics and likelihood ratio framework for reporting the evidential value of spectra, Anal. Chim. Acta 931 (2016) 34–46, https://doi.org/10.1016/j.aca.2016.05.016.
[16] X.-h. Chen, C. Champod, X. Yang, S.-p. Shi, Y.-w. Luo, N. Wang, Y.-c. Wang, Q.-m. Lu, Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features, Forensic Sci. Int. 282 (2018) 101–110, https://doi.org/10.1016/j.forsciint.2017.11.022.
[17] A.J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, I. Alberink, Performance study of a score-based likelihood ratio system for forensic fingermark comparison, J. Forensic Sci. 62 (3) (2017) 626–640, https://doi.org/10.1111/1556-4029.13339.
[18] N. Brümmer, J. Du Preez, Application-independent evaluation of speaker detection, Comput. Speech Lang. 20 (2–3) (2006) 230–275, https://doi.org/10.1016/j.csl.2005.08.001.
[19] T. Kenter, M. De Rijke, Short text similarity with word embeddings, Proceedings of the 24th ACM international on conference on information and knowledge management, 2015, pp. 1411–1420.
[20] L. Jeancolas, D. Petrovska-Delacrétaz, G. Mangone, B.-E. Benkelfat, J.-C. Corvol, M. Vidailhet, S. Lehéricy, H. Benali, X-vectors: new quantitative biomarkers for early parkinson's disease detection from speech, Front. Neuroinform. 15 (2021) 578369, https://doi.org/10.3389/fninf.2021.578369.
[21] A.R. Naini, A. Rao, P.K. Ghosh, Whisper to neutral mapping using cosine similarity maximization in i-vector space for speaker verification, Interspeech (2019) 4340–4344.
[22] X. Zhang, R. Wu, Z. Qu, A cosine-similarity-based deconvolution method for analyzing data-independent acquisition mass spectrometry data, Appl. Sci. 13 (10) (2023) 5969, https://doi.org/10.3390/app13105969.
[23] T.V. Harwood, D.G. Treen, M. Wang, W. de Jong, T.R. Northen, B.P. Bowen, BLINK enables ultrafast tandem mass spectrometry cosine similarity scoring, Sci. Rep. 13 (1) (2023) 13462, https://doi.org/10.1038/s41598-023-40496-9.
[24] D. van der Vloed, Data strategies in forensic automatic speaker comparison, Forensic Sci. Int. 350 (2023) 111790, https://doi.org/10.1016/j.forsciint.2023.111790.
[25] D. Ramos, R. Haraksim, D. Meuwly, Likelihood ratio data to report the validation of a forensic fingerprint evaluation method, Data Brief 10 (2017) 75–92, https://doi.org/10.1016/j.dib.2016.11.008.
[26] J. Malmborg, L. Joborn, M. Beming, A. Nordgaard, I. Alberink, Comparing a machine learning approach with traditional methods for forensic source attribution using chromatographic data, Forensic, Chemistry (2025) 100699, https://doi.org/10.1016/j.forc.2025.100699.
[27] T. Korpinsalo, J. Rautavirta, S. Huhtala, T. Reinikainen, J. Corander, Forensic comparison of amphetamine chemical profiles by bayesian predictive modelling, J. Chemom. 38 (12) (2024) e3630, https://doi.org/10.1002/cem.3630.
[28] G. Sauzier, W. van Bronswijk, S.W. Lewis, Chemometrics in forensic science: approaches and applications, Analyst 146 (8) (2021) 2415–2448, https://doi.org/10.1039/D1AN00082A.
[29] M. Diez, A. Varona, M. Penagarikano, L.J. Rodriguez-Fuentes, G. Bordel, Optimizing pllr features for spoken language recognition, 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 779–784.
[30] K.A. Martire, R.I. Kemp, M. Sayle, B.R. Newell, On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect, Forensic Sci. Int. 240 (2014) 61–68, https://doi.org/10.1016/j.forsciint.2014.04.005.
[31] G.J. Nohynek, R. Fautz, F. Benech-Kieffer, H. Toutain, Toxicity and human health risk of hair dyes, Food Chem. Toxicol. 42 (4) (2004) 517–543, https://doi.org/10.1016/j.fct.2003.11.003.
[32] K.L. Monson, J.R. Brown, CODIS: a national index of DNA identification records, advances in forensic haemogenetics: 15th congress of the international society for forensic haemogenetics (internationale gesellschaft für forensische hämogenetik eV), venezia, 13–15, Springer 1994 (1993) 286–288.
[33] C. Milton, Combined DNA Index System (CODIS), Cold Case Homicides, CRC Press, 2017, pp. 365–370.