Breaking with trends in forensic dating: A likelihood ratio-based comparison approach

Forensic Science International

Breaking with trends in forensic dating: A likelihood ratio-based comparison approach

--Manuscript Draft--

Manuscript Number:	FSI-D-23-00159R2
Article Type:	Original Research Article
Keywords:	bloodstains; Time since deposition, Forensic dating, Likelihood ratio, Comparison problem
Corresponding Author:	Alicja Menżyk University of Silesia Katowice, POLAND
First Author:	Alicja Menżyk
Order of Authors:	Alicja Menżyk
	Agnieszka Martyna
	Alessandro Damin
	Marco Vincenti
	Grzegorz Zadora
Abstract:	Further steps toward understanding the time-related information contained within bloodstains found at the crime scene are rightly considered a top priority in forensic science. Contrary to widely held assumptions, the reason for the delayed exploitation of bloodstains dating methods in practice is not the lack of suitable analytical techniques for monitoring degradation processes. The problem lies in the variability of the environmental and circumstantial conditions, playing a vital role in the degradation kinetics of blood deposits. The present article demonstrates the possibility of breaking with current approaches based on absolute age estimations to finally answer time-centered questions in real forensic scenarios. The proposed novel framework for situating forensic traces in time is based on the likelihood ratio assessment of the (dis)similarity between the evidence decomposition and sets of reference materials obtained through supervised aging. In such a strategy, every dating procedure is constructed on a case-by-case basis to fit examined blood traces, thereby limiting the adverse influence of external factors on the validity of age estimations and providing a way for future crime scene implementation.
Suggested Reviewers:	Theresa Stotesbury theresa.stotesbury@ontariotechu.ca
	Paolo Oliveri paolo.oliveri@unige.it

Breaking with trends in forensic dating: A likelihood ratio-based comparison approach*

Alicja Menżyk,[a,b] Agnieszka Martyna,[a] Alessandro Damin,[c] Marco Vincenti[c, d] and Grzegorz Zadora[a,b]

[a] Forensic Chemistry Research Group, Institute of Chemistry, University of Silesia in Katowice, Szkolna 9, 40-006 Katowice, Poland

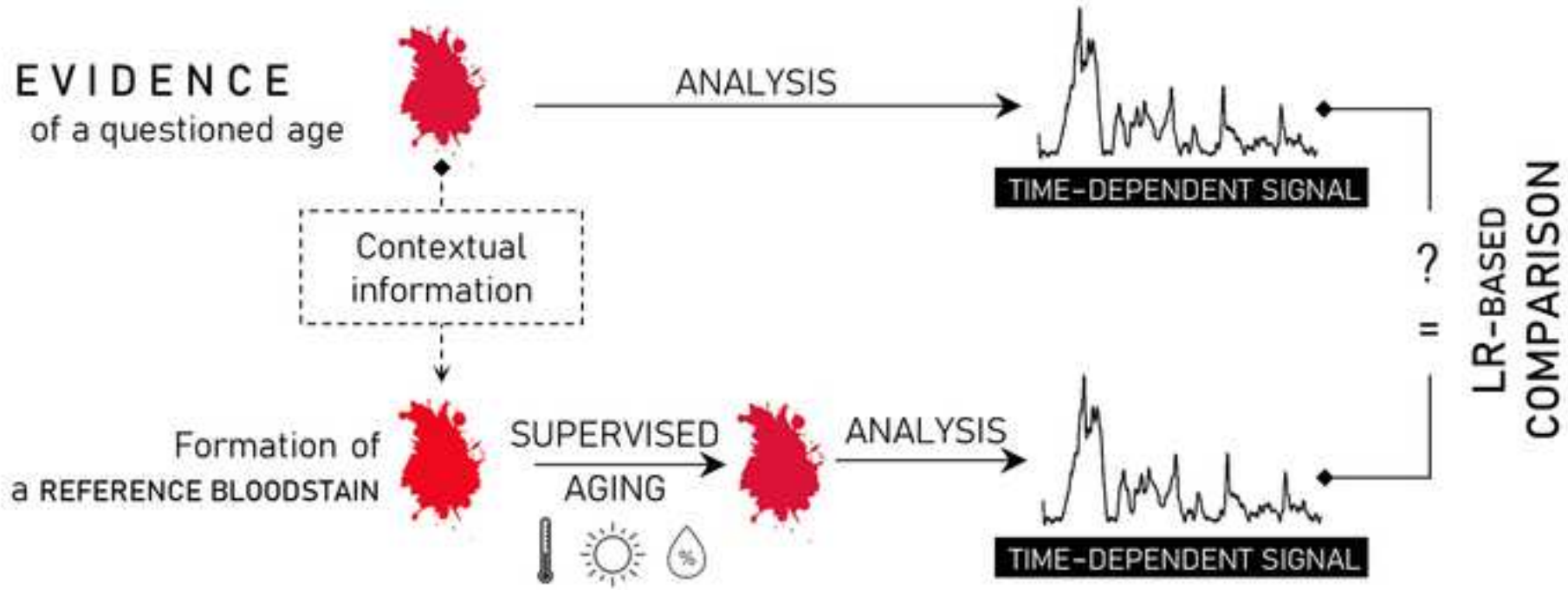
E-mail: alicja.menzyk@us.edu.pl

[b] Institute of Forensic Research in Krakow, Westerplatte 9, 31-003, Krakow, Poland

[c] Dipartimento di Chimica, Universita degli Studi di Torino, Via P. Giuria 7, 10125 Torino, Italy

[d] Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Regione Gonzole 10/1, 10043 Orbassano, Torino, Italy

* In memory of Professor Gianmario Martra: Teacher, Colleague and Friend.



Highlights

- The novel framework for forensic dating of bloodstains is established
- The conventional dating approach is substituted with the comparison problem
- The similarity between evidence and reference materials degradation is assessed
- Likelihood ratio models are developed for solving the comparison problem

Breaking with trends in forensic dating: A likelihood ratio-based comparison approach

Supplementary Material available.

Abstract

Further steps toward understanding the time-related information contained within bloodstains found at the crime scene are rightly considered a top priority in forensic science. Contrary to widely held assumptions, the reason for the delayed exploitation of bloodstains dating methods in practice is not the lack of suitable analytical techniques for monitoring degradation processes. The problem lies in the variability of the environmental and circumstantial conditions, playing a vital role in the degradation kinetics of blood deposits. The present article demonstrates the possibility of breaking with current approaches based on absolute age estimations to finally answer time-centered questions in real forensic scenarios. The proposed novel framework for situating forensic traces in time is based on the likelihood ratio assessment of the (dis)similarity between the evidence decomposition and sets of reference materials obtained through supervised aging. In such a strategy, every dating procedure is constructed on a case-by-case basis to fit examined blood traces, thereby limiting the adverse influence of external factors on the validity of age estimations and providing a way for future crime scene implementation.

Keywords: Bloodstains; Time since deposition, Forensic dating, Likelihood ratio, Comparison problem

1. Introduction

Forensic experts – nowadays equipped with DNA profiling methods – have never been better prepared to answer questions about a suspect's identity, which usually arise during the investigation process. Genetic typing of biological evidence – such as bloodstains – has clearly revolutionized forensic science to such an extent that it is now one of the cornerstones of modern policing. However, in the era of DNA testing, it may be overlooked that identifying the donor of the collected biological trace is not always the most critical issue. Thus, to demonstrate a possibly strong link between the evidence and the investigated crime, it is often necessary to prove that the questioned material was created during the incriminating event [1]. This can be achieved by providing information about the time elapsed since trace deposition (TSD, time since deposition).

A reliable answer to the question of bloodstains' age could be a significant added value to the criminal inquiry, allowing the judicature to recreate a criminal situation *a posteriori* [2, 3]. Information about the time of bloodstain(s) formation may prove helpful directly during the investigative phase of the criminal procedure, for example, by establishing a timeline of events through sequencing the deposition of traces, but not only. The time-centered questions often call for rational responses also at the evaluative stage of the proceeding when hypotheses representing the prosecutor's and defense's standpoints are confronted. At this phase, a common defense tactic is to question the validity of evidential bloodstains deposited by the suspect at the crime scene. In such a case, the prosecutor's opponents do not challenge the identification of blood traces but rather the time of their formation, thereby casting doubts on the suspect's presence at the crime scene during the incriminating event. In this context, the role of forensic experts is to indicate which hypotheses presented by both parties to the proceedings are supported by the evidence, namely the age of the questioned blood traces. Consequently, correct estimation of TSD could help verify bloodstains' relevance – and hence also the suspect's connection – to the case considered in the courtroom scenario.

Awareness of the gravity of this temporal information – or lack thereof – has accompanied forensic experts and judiciary members for at least a hundred years [4]. Nevertheless, despite significant research efforts [5–7], a reliable method for bloodstains dating is still unavailable. The root cause of the problem is not the researchers' inability to develop an analytical tool capable of monitoring the time-dependent properties of bloodstains, as evidenced by a constantly growing body of research in forensic dating [5–7]. Indeed, spectroscopic methods, particularly Raman spectroscopy, have contributed to a better understanding of degradation mechanisms inside blood traces [6, 8–15]. However, despite these remarkable technological advancements, none of the dating approaches has yet to overcome the experimental research phase. No forensic organization working according to ISO/IEC 17025 international quality standards performs the absolute dating of bloodstains on a daily analytical basis [16]. This proves that obtaining sound analytical results does not accurately address the entire problem,

as the mere chemical information about the aging process of blood – no matter how detailed – turns out to be insufficient if not integrated into an appropriate dating framework.

A closer inspection of previous research suggests that the delayed exploitation of the developed dating models in forensic practice might depend on the data analysis approach. According to the commonly adopted strategy (hereinafter referred to as the *conventional dating approach*), most proposed methods have sought the dependency between the TSD and some dynamic properties of blood, degrading under stable, strictly controlled conditions through (multivariate) regression analysis [5, 6]. However, apart from yielding dating models of excellent performance in laboratory settings, the conservative practice of controlling factors affecting the aging kinetics (e.g., environmental conditions) may pose certain risks. These models – consistently fitting the aging behavior of training samples used for their construction – are not equally effective in the typical multivariate and hardly predictable forensic practice. In other words, forensic experts are faced here with the age-old question of the transferability of proposed conventional dating methods, undoubtedly powerful in artificial laboratory settings, to more realistic contexts.

External factors will always affect the blood degradation process, altering its rate to a lesser or greater extent. This means that when dealing with evidence, the uncertainty deriving from these influencing agents cannot be eliminated or universally estimated in advance. It should be incorporated into the dating strategy. Such a shift in perspective is proposed in the present work. According to the strategy adopted in this study, verifying the relevance of blood evidence to the considered crime should not necessarily be performed using previously established calibration models. A strong dependency between the accuracy of TSD estimations and the aging kinetics of blood at the crime scene constitutes a major contraindication to adopting this conventional approach. It is hypothesized instead that in order to situate the trace in time, it suffices to compare the evidential material with reference bloodstains, created in such a way as to match the local conditions and time since deposition (TSD) corresponding to the prosecution's/defense's version of events (e.g., t_R in Fig. 1).

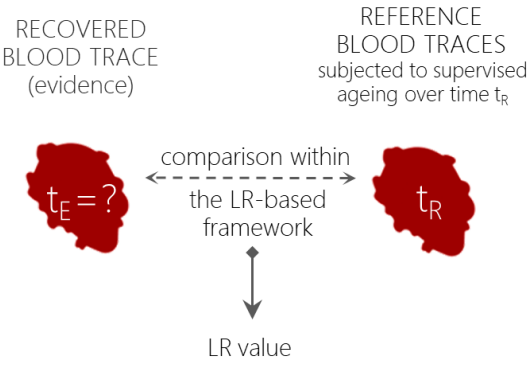


Fig. 1 Implementation of the comparison procedure within the LR framework to estimate the TSD of a questioned bloodstain.

The critical stage of this comparison is the likelihood ratio-based (LR-based) assessment [17–19] of the (dis)similarity between the analytical signals reflecting the decomposition stage of evidence

and reference materials obtained through supervised aging. This process is designed to recreate the degradation of the evidence on site as precisely as possible to normalize the aging kinetics of compared materials (evidence and reference bloodstains). It means that every dating procedure would be constructed case-by-case, each time tailored to fit the examined traces. In such a way, the influence of external factors (e.g., environmental conditions) on the aging kinetics and, through that, on the analysis's validity should be considerably reduced, providing a way for future crime scene implementation.

2. Materials and methods

2.1 Bloodstains databases

Since the main idea behind the novel dating framework developed herein is the LR-based comparison of recovered (evidence) and reference bloodstains (samples of known age), at least two databases were required for the practical proof of concept. One set was the source of data for training the LR model and deriving the reference sample of bloodstain (the product of the so-called supervised aging process) and the second set from which the recovered bloodstain (acting as the evidence) was selected. Eventually, two separate databases (hereafter referred to as sets No. 1 and 2) were prepared at a three-weeks time distance, according to the scheme presented in Fig. 2.

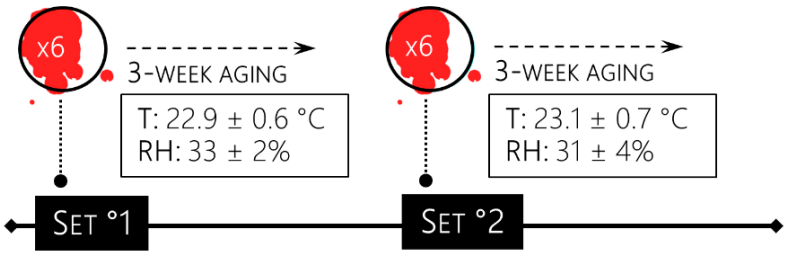


Fig. 2 The general scheme of the supervised aging process of two different bloodstain sets served to develop and validate LR models.

Each database consisted of spectral signatures of degrading blood traces created by depositing 20 µL of additive-free capillary blood (derived from a single donor) in aluminum sample pans. During the first day of degradation, blood deposits (six per database) were analyzed for up to eight hours elapsed since bloodstain formation with a two-hour interval. Due to the significant slow-down in the degradation process [13], the time resolution of measurements was subsequently reduced, and spectra were registered daily (excluding weekends) for the next three weeks. This measurement protocol resulted in databases consisting of spectra corresponding to 18 time-points, assumed to constitute 18 different evidence time-related sources (*source* should be understood as a group of bloodstains corresponding to the same TSD). Each time-point was characterized by six bloodstains (which served as replicate measurements of samples at a given time-point) created one after another at appropriate time intervals, estimated with regard to the duration of spectral acquisition. In order to meet the criteria of supervised aging, between measurements, all samples were stored in darkness, exposed to temperature (T) and

relative humidity (RH) controlled by the laboratory air-conditioning system to normalize the degradation kinetics between each dataset as much as possible (Fig. 2).

It should also be added that within this work scope, Raman spectroscopy (RS) was selected to characterize the state of bloodstains degradation. The rationale for this choice was given in our previous work [13] that demonstrated the capability of RS to deliver information inherent to chemical changes accompanying the *ex vivo* aging of blood. Raman signatures were obtained using a Renishaw *inVia Raman Microscope,* following the procedure developed in [13]. To prevent sample damage from the excessive point laser irradiation and increase the representativeness of a single measurement, all spectra were registered during the magnetic-driven rotation of the sample [20, 21]. Detailed experimental conditions used are summarized in Table 1.

Table 1 The rotating mode experimental conditions used during the spectral characterization of bloodstains.

Experimental conditions	
Laser excitation source [nm]	785
Laser density [mW/μm^2]	ca. 0.16 (10% of total laser power)
Spectral range [cm^{-1}]	600–1800
Objective	5× NIR optimized objective (N.A. = 0.1)
Accumulations	2
Time of acquisition [s]	20

2.2 The likelihood ratio procedure for bloodstains comparison

In judicial practice, verifying the association between the evidence and reference materials to establish whether they could originate from the same (H_1) or two different sources (H_2) is defined as a classical comparison problem [17–19]. In this study, however, we are dealing with its unusual variation (depicted in Fig. 1), where the *source* should be understood as a group of bloodstains corresponding to the same TSD. Thus, the considered hypotheses can be formulated as follows:

H_1: the age of the recovered evidence (questioned bloodstain(s) revealed at the crime scene; t_E) MATCHES the age of the reference material (bloodstains obtained during the process of supervised aging according to the prosecutor's or defense standpoint; t_R): $t_E = t_R$.

H_2: the age of the recovered evidence (questioned bloodstain(s) revealed at the crime scene; t_E) DOES NOT MATCH the age of the reference material (bloodstains obtained during the process of supervised aging according to the prosecutor's or defense standpoint; t_R): $t_E \neq t_R$.

Even though such a comparison implies a discrimination task, easily solvable with chemometric tools, in actuality, establishing the (dis)similarity of bloodstains by comparing the likeness of the data solely (e.g., Raman spectra) is insufficient. It should be reminded that adequately performed forensic expertise also involves interpreting the findings to establish their evidential value. For the forensic assessment of evidence within a comparison problem, it is essential not only the similarity but also the

rarity of the compared patterns in the relevant population. Thus, emphasis has to be put on the question: what does observed similarity mean in this particular case?

To answer that question, the likelihood ratio (LR) framework [17–19] is recommended, which considers not only the similarity and the rarity of physicochemical characteristics but also the possible levels of variation (within- and between-source variability). The LR quantifies the strength of sources' similarity, which escalates with their increasing rarity, indicating how strongly they are alike to verify whether they share a common source. It is computed as the probability of observing the physicochemical data for the samples (E), given the propositions (H_1 and H_2):

$$\boldsymbol{LR} = \frac{\mathbf{Pr}(\boldsymbol{E}|\boldsymbol{H_1})}{\mathbf{Pr}(\boldsymbol{E}|\boldsymbol{H_2})}.$$

The LR can have a value between zero and infinity, and its interpretation is the following: LR values above one support H_1, while LR values below one provide support for H_2. Moreover, the higher (lower) the LR value, the stronger the H_1 (H_2) support. Therefore, LR is a qualitative and quantitative measure of the strength of the support for one of the hypotheses, which – in the forensic context – is a highly desirable feature.

The conventional LR approach, however, is not well-adjusted to deal with high-dimensionality data [17] where the number of variables (*J*) exceeds the sample population's size (*I*). When $J > I$, the LR models fail due to the singularity or the instability of the inverse of some variance-covariance matrices; thus, reducing data dimensionality becomes obligatory. A remedy for this obstacle is a form of symbiosis between the LR approach and chemometrics known as hybrid LR models, introduced by Martyna *et al.* [22], which have proven their effectiveness through years of use in forensic studies [23–26]. Reduction of data dimensionality and extraction of informative features (concerning the purpose of the study) are carried out through the implementation of chemometrics (*2.2.1 The chemometric tactic of data treatment*), resulting in the subsequent construction of LR models based on a limited number of new latent variables (*2.2.2 The development and validation of hybrid likelihood ratio models*).

2.2.1 The chemometric tactic of data treatment

Well-performing LR models for discrimination purposes are defined for just a few latent variables that maximize the between-source variation, $\boldsymbol{B}$, while minimizing the data variation within each source, $\boldsymbol{W}$. One method that allows finding these new latent variables is rMANOVA [27], whose task is to find the directions along which the between-source variance is the highest and the within-source variance is the lowest. These directions are defined as the eigenvectors of the matrix

$$[(1-\delta)W + \delta T]^{-1}B.$$

$\boldsymbol{T}$ is the target matrix, which is either $\boldsymbol{T} = \frac{\mathbf{1}}{\boldsymbol{p}}\boldsymbol{tr}(\boldsymbol{W})$ or $\boldsymbol{T} = \boldsymbol{diag}(\boldsymbol{W})$ when the variances of *p* variables for each source are equal or unique, respectively. In these studies, equal variances were assumed for each source to keep consistent with the LR model's assumption. The δ estimation depends

on the chosen target and evaluates the variance of the $\boldsymbol{W}$ components according to the Ledoit-Wolf theorem [27]. It should also become clear that the choice of rMANOVA was not accidental, as it demonstrates the possibility of solving two problems simultaneously. It reduces data dimensionality in a beneficial way for optimal source separation and highlights their specific features. For the above reasons, the LR models proposed for comparative dating of bloodstains were built for the first latent variable, LV1, from rMANOVA, which in our recent studies [27] proved reasonably effective in the separation of the sources described by highly multivariate data.

rMANOVA was performed for the signals subjected to adequate transformations to eliminate undesired spectral contributions – undermining variation of interest – and expose the vital features for further analysis. These mathematical transformations were effected through a suitable combination of preprocessing techniques applied to the raw signals, which – in the case of Raman spectra – usually includes denoising/smoothing, baseline correction, and normalization. In this research, the preprocessing aimed to support the LR models' performance by uniquely defining the characteristics of each source referred to different time points of bloodstains deposition. For this reason, a concept based on genetic algorithms (GA) previously developed by A. Martyna et al. [28] was implemented that remarkably supports the discrimination analysis of the signals owing to a case-shaped optimization of the preprocessing strategy. To measure the performance of the given combination of methods, reflected in the effectiveness of sources (differently-aged bloodstains) separation, a quality parameter (being also a fitness function) was defined as a ratio of the between-source and within-source variation (b^2/w^2) of the first latent variable (LV1) from the regularised MANOVA (rMANOVA) [27].

The applied preprocessing strategy was initialized with denoising/smoothing, followed by baseline correction and normalization. Additionally, to compensate for heteroscedastic noise [29] that grows with signal intensity and thus affects the covariance structure of the variables, denoised/smoothed signals, $\boldsymbol{a} = [\boldsymbol{a_1}, \ldots, \boldsymbol{a_J}]$ were subjected to the centered log-ratio transform (CLR) in the following manner:

$$\boldsymbol{s} = \boldsymbol{log_{10}a} - \boldsymbol{1/J}\sum_{\boldsymbol{i=1}}^{\boldsymbol{J}} \boldsymbol{log_{10}a_i}.$$

The most optimal preprocessing strategy was indicated using a GA-based procedure from among chromosomes, each described by three genes referring to denoising, baseline correction, and normalization steps. The tested denoising strategies involved 16 different methods based on discrete wavelet transform (DWT) [30] and Savitzky-Golay (SG) filter [31]. The baseline correction methods summed up to 64 various tools using asymmetric penalized least squares (AsPLS) [32–36], robust baseline estimation (RBE) [37], statistics-sensitive non-linear iterative peak-clipping (SNIP) [38], polynomials fitting [39, 40], and quantile regression (polyQR) [41]. Finally, standard normal variate (SNV) and probabilistic quotient normalization (PQN) [42, 43] were introduced among the available options of normalization methods. The groups of parameters, which characterize the considered preprocessing techniques, were roughly selected for optimization after a visual examination of the

preprocessed signals. Table S1 summarises all preprocessing methods implemented in the GA optimization procedure.

The initial generation in GA entailed 50 randomly selected preprocessing strategies. The chance of mutations was 0.1, the elitism level was 5%, and the algorithm converged when three subsequent solutions were identical.

2.2.2 The development and validation of hybrid likelihood ratio models

The objective of developed LR models is to compare the characteristics of the reference material from the first database (set No. 1) and the potential evidence – the recovered sample originating from set No. 2 – to assess whether they can be of the same (H_1) or different (H_2) age. As a reminder, considered hypotheses were defined in *2.2 The likelihood ratio procedure for bloodstains comparison*. In the beginning, the following procedure was proposed for comparing two bloodstains. Set No. 1 and 2 were subjected to the same preprocessing procedure, selected using the GA for the set, which collects the reference samples. Table 2 lists the selected preprocessing strategies, depending on whether the reference samples used for training the model were part of set No. 1 or 2. The first database (set No. 1), containing six Raman spectra recorded for 18 time-related bloodstains (6×18 signals), was divided into two parts. Three signals corresponding to 18 time-related points (3×18 signals) were randomly selected to constitute a sub-database (set No. 1a). The remaining created the second equally-sized sub-database (set No. 1b). Set No. 1a was used for summarising the case-specific data (providing information about bloodstains within each time-point) to find the new LV1 direction from rMANOVA, which best separates the spectra from different time-related points. The data from set No. 1b and the spectra measured for the recovered bloodstains (constituting set No. 2) were then mean-centered (using the mean of set No. 1a) and projected on LV1. Finally, the projections for the recovered sample, defined by the mean of $k_1 = 3$ data $\bar{\boldsymbol{y}}_{\boldsymbol{1}} = \sum_{i=1}^{k_1} \boldsymbol{y}_{1i}/\boldsymbol{k}_{\boldsymbol{1}}$ and one of the reference bloodstains, defined by a mean of $k_2 = 3$ data $\bar{\boldsymbol{y}}_{\boldsymbol{2}} = \sum_{i=1}^{k_2} \boldsymbol{y}_{2i}/\boldsymbol{k}_{\boldsymbol{2}}$ were compared. To define the relevant distributions, the remaining $M = 17$ samples with $n = 3$ data each, described by their means $\bar{\boldsymbol{x}}_{\boldsymbol{m}} = \sum_{j=1}^{n} \boldsymbol{x}_{mj}/\boldsymbol{n}$, were used as a background population.

The within-source variance was defined as:

$$\boldsymbol{w^2} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{n}(\boldsymbol{x}_{mj} - \bar{\boldsymbol{x}}_{\boldsymbol{m}})^2}{\boldsymbol{M(n-1)}},$$

while the between-source variance was expressed as:

$$\boldsymbol{b^2} = \frac{\sum_{m=1}^{M}(\bar{x}_m - \bar{x})^2}{\boldsymbol{M(n-1)}} - \frac{\boldsymbol{w^2}}{\boldsymbol{n}},$$

where $\bar{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{Mn}}\sum_{m=1}^{M}\sum_{j=1}^{n}\boldsymbol{x}_{mj}$.

The LR formula applied herein consists of four components, as shown in Fig. 3. Its first part (denoted in Fig. 3 as LR_1) expresses the probability density of the distribution with a mean equal to zero

and variance equal to $\frac{w^2}{k_1}+\frac{w^2}{k_2}$ at $\bar{y}_1-\bar{y}_2$. The second component (LR_2) estimates the probability density of the distribution estimated with Gaussian kernels positioned at each $\bar{x}_m$ with variance equal to $\frac{w^2}{k_1+k_2}+h^2b^2$ at $\bar{y}^*=\frac{\bar{y}_1k_1-\bar{y}_2k_2}{k_1+k_2}$. The h is expressed as $\left(\frac{4}{M(2p+1)}\right)^{\frac{1}{p+4}}$ where p is the number of the considered variables. In the case of this study, the p equals one. The final density curve is obtained by averaging Gaussian kernels to integrate into the unit area [44].

In contrast to the first two components, which assume that H_1 is true, the remaining two expressions in the denominator of the LR formula assume the opposite, so they are independent. They both estimate the probability density of the distribution estimated with Gaussian kernels positioned at each $\bar{x}_m$ with variance equal to $\frac{w^2}{k_1}+h^2b^2$ (or $\frac{w^2}{k_2}+h^2b^2$) at $\bar{y}_1$ (LR_3) or $\bar{y}_2$ (LR_4).

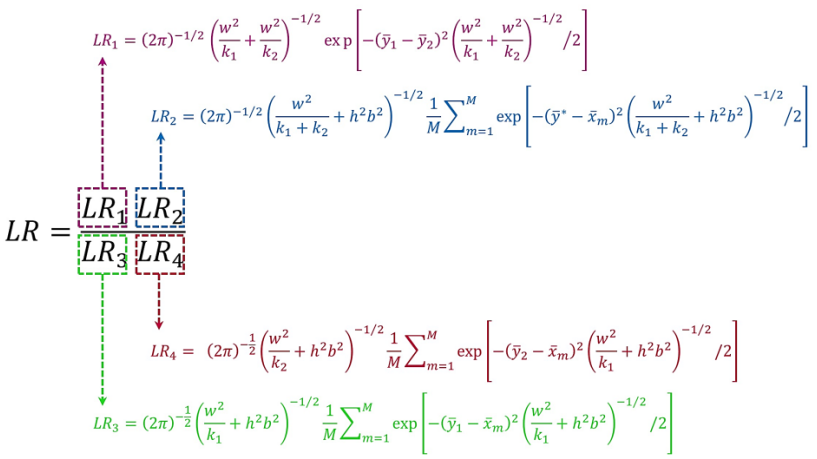


Fig. 3 The expression for the likelihood ratio (LR) applied in the study.

Subsequently, this division was reversed; the second database (set No. 2) was used to train the models, and the first database (set No. 1) provided the evidence samples for the comparison procedure. Models were constructed based on seven different ranges of the Raman spectrum (here, denoted as A–G; for details, see Fig. S1 or Table 2), resulting in 14 different LR models. These spectral ranges were recognized in our previous article [13] as highly informative (from the dating perspective) thanks to regularised MANOVA (rMANOVA). They consisted of time-dependent features characterized by b^2/w^2 above one, meaning that the variance between differently-aged bloodstains (b^2) was higher than within groups of bloodstains characterized by the same age (w^2).

Having established LR models, their validation was undertaken. For research purposes, such a procedure in which each G bloodstain samples from the first database (set No. 1) is compared with a single recovered sample from the second database was insufficient. Therefore, to validate the LR model, the procedure was repeated for G recovered samples stored in the second database (set No. 2). In this way, there were G^2 comparisons, of which G concerned comparing samples of the same age (i.e., $t_R = t_E$). Since the H_1 hypothesis should be supported in this case, and the expected LR value is above one,

false negatives appear when LR < 1. To estimate the false positive answers, where the H_2 should be supported (LR < 1), $\boldsymbol{G^2 - G}$ concerned comparing samples of different ages (i.e., $t_R \neq t_E$); thus, each LR > 1 is classified as the answer providing incorrect support for the H_1 (false positive). Both databases (sets 1 and 2) were used to test the performance of the LR models. Moreover, the false positives and negatives experiments were carried out ten times, each time generating a set to train and test the LR model. Therefore, the final error rates will be presented as an average of these partial results.

Additionally, the quantitative aspect of LR models' functioning, namely their ability to point out the strength of support for each hypothesis, was assessed using empirical cross entropy (ECE) [17, 45, 46]. The ECE is a performance metric that relies on assigning an appropriate penalty for each yielded LR value. The penalty magnifies with the increasing support for the incorrect hypothesis according to the logarithmic strictly proper scoring rules (a description of the ECE procedure for the performance assessment is presented in the Supplementary Material, SM). Consequently, the higher the support for the wrong hypothesis, the more severe the given error is because the greater penalty is assigned to the evaluated LR model.

All calculations were performed in R software [47] using scripts home-written by A. Martyna and available R packages.

3. Results and discussion

The following example may be considered to understand the comparison approach proposed in this study. Among the traces secured at the crime scene, there were bloodstains identified as originating from the suspect. The defense questioned the relevance of these findings by casting doubt on the timing of the bloodstains' deposition. Thus, the critical information for establishing the suspect's connection to the considered event was the time of the trace formation. Contrary to the conventional dating approach, obtaining this information does not have to come down to determining the absolute age of the sample by applying a previously established calibration model. It suffices to compare the degradation state of the evidential material with some reference bloodstains, created in such a way to match hypotheses, which represent the position of the prosecutor and the defense.

To obtain a well-balanced evaluation of the evidence, it would be desirable to create two different sets of reference materials, producing – at best – results of two different comparisons. These reference materials are obtained through supervised aging, simulating – as closely as possible – the actual settings of evidence decomposition at the crime scene. In other words, a similar volume of blood, freshly drawn from the donor (the same individual who was the donor of the questioned trace), should be deposited on an identical substrate and stored under similar conditions as those observed at the crime scene. If possible, one of these reference sets is then subjected to the degradation process for the time corresponding to the prosecution's version of events. The other should be aged according to the defense's scenario.

Having both evidential and reference materials, the proposed dating procedure enters the next phase, namely the characterization of the state of blood traces degradation. Finally, the resulting physicochemical data (e.g., spectral signatures) should be interpreted in the context of the prosecutor's and the defense's perspectives and communicated understandably to the justice system representatives. An approach founded on the likelihood ratio framework [16–18] that forms the proposed comparative procedure's backbone is recommended to achieve this goal.

3.1. Preprocessing of Raman signatures using genetic algorithm (GA)

As well-documented, the correct choice of preprocessing usually improves the performance of statistical models, including LR models, which – at the same time – can be significantly impaired in case of improperly handled signals. Altogether, adequate preprocessing is a crucial step in data treatment. In the case of Raman spectroscopy-derived signals, usually, two basic preprocessing steps are vitally crucial for feasible analysis. The first is baseline correction methods to remove the baseline effects arising, among others, from fluorescence and other additive features in the spectra. The second group consists of normalization procedures whose task is to remove multiplicative effects related, for instance, to laser intensity fluctuations or out-of-focus contributions. Thus, normalization usually boils down to multiplying the signal by a scaling value to make the corresponding intensities, which, in theory, should not pose any differences as comparable across spectra as possible. Unfortunately, regardless of the final goal of data modeling, the difficulty here is that there is no universal combination of preprocessing methods suitable for each specific analysis's objective and the type of spectral distortions.

In order to address this problem, before developing LR models, the strategy based on applying the genetic algorithm (GA) to indicate the combination of preprocessing tools – optimal given the study purpose – was implemented. Table 2 collects all the preprocessing solutions suggested by the GA as the most effective combinations in yielding the highest between-source and within-source variance (b^2/w^2) ratio for the first latent variable computed from rMANOVA as a quality parameter. In such a way, as broadly discussed in the study of Martyna [28], the discrimination of signals corresponding to differently-aged bloodstains was significantly enhanced, as the preprocessing procedure ensured the best exposition of differences between sources while minimizing the casual variations within sources. The suitability of the proposed methodology was ultimately verified by the performance of developed LR models built for the first rMANOVA's eigenvector, which were reflected in – among others – the levels of false positive and false negative rates (shown as boxplots in Fig. 4) and also numerically summarized in Table S2.

Table 2 Details of preprocessing strategies applied when sets No. 1 and No. 2 were training sets found using a genetic algorithm (GA). All abbreviated methods are explained in Table S1 (SM) and partially in section *2.2.1*.

Spectral range [cm^{-1}]	Reference: set No. 1	Reference: set No. 2
A: 1200-1410	**Denoising**: SG, polynomial degree p = 3 **Baseline correction**: SNIP, clipping window w = 26 **Normalization**: PQN	**Denoising**: DWT, Daubechies Least Asymmetric 4, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: RBE, h = 0.4, b = 2 **Normalizatio**n: PQN
B: 1200-1300; 1500-1700	**Denoising**: DWT, Daubechies Least Asymmetric 8, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: CWTAsWPLS, m = 2, $\lambda = 1 \cdot 10^8$ **Normalization**: SNV	**Denoising**: DWT, Coiflets 1, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: polyQR, polynomial degree p = 5, q = 0.05 **Normalization**: SNV
C: 1200-1300; 950-1020	**Denoising**: DWT, Coiflets 5, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: RBE, h = 0.4, b = 2.2 **Normalization**: PQN	**Denoising**: DWT, Daubechies Least Asymmetric 4, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: RBE, h = 0.3, b = 2.5 **Normalization**: PQN
D: 1200-1410; 950-1020	**Denoising**: DWT, Coiflets 5, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: multiWAsPLS, m = 2, λ = 10, $\mu = 1 \cdot 10^9$ **Normalization**: SNV	**Denoising**: DWT, Coiflets 1, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: polyQR, polynomial degree p = 5, q = 0.05 **Normalization**: PQN
E: 1200-1410; 1500-1700	**Denoising**: SG, polynomial degree p = 3 **Baseline correction**: polyQR, polynomial degree p = 6, q = 0.01 **Normalization**: SNV	**Denoising**: SG, polynomial degree p = 5 **Baseline correction** pWAsPLS, m = 2, $\lambda = 1 \cdot 10^6$, w = 0.001 **Normalization**: SNV
F: 1200-1300; 1500-1700; 950-1020;	**Denoising**: DWT, Daubechies Least Asymmetric 4, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: CWTAsWPLS, m = 2, $\lambda = 2 \cdot 10^8$ **Normalization**: SNV	**Denoising**: DWT, Daubechies Least Asymmetric 8, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: SNIP, clipping window w = 30 **Normalization**: SNV
G: 1200-1410; 1500-1700; 950-1020;	**Denoising**: SG, polynomial degree p = 4 **Baseline correction**: SNIP, clipping window w = 26 **Normalization**: SNV	**Denoising**: DWT, Coiflets 1, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: multiWAsPLS, m = 2, λ = 10, $\mu = 1 \cdot 10^8$ **Normalization**: SNV

d – decomposition level for denoising, t – threshold estimation, c – thresholding policy, sd – dispersion estimate, m – order of differences, λ – penalty, μ – penalty term, h – the proportion of signal points for the local regression, b – robustness parameter, q – quantile, w – weights

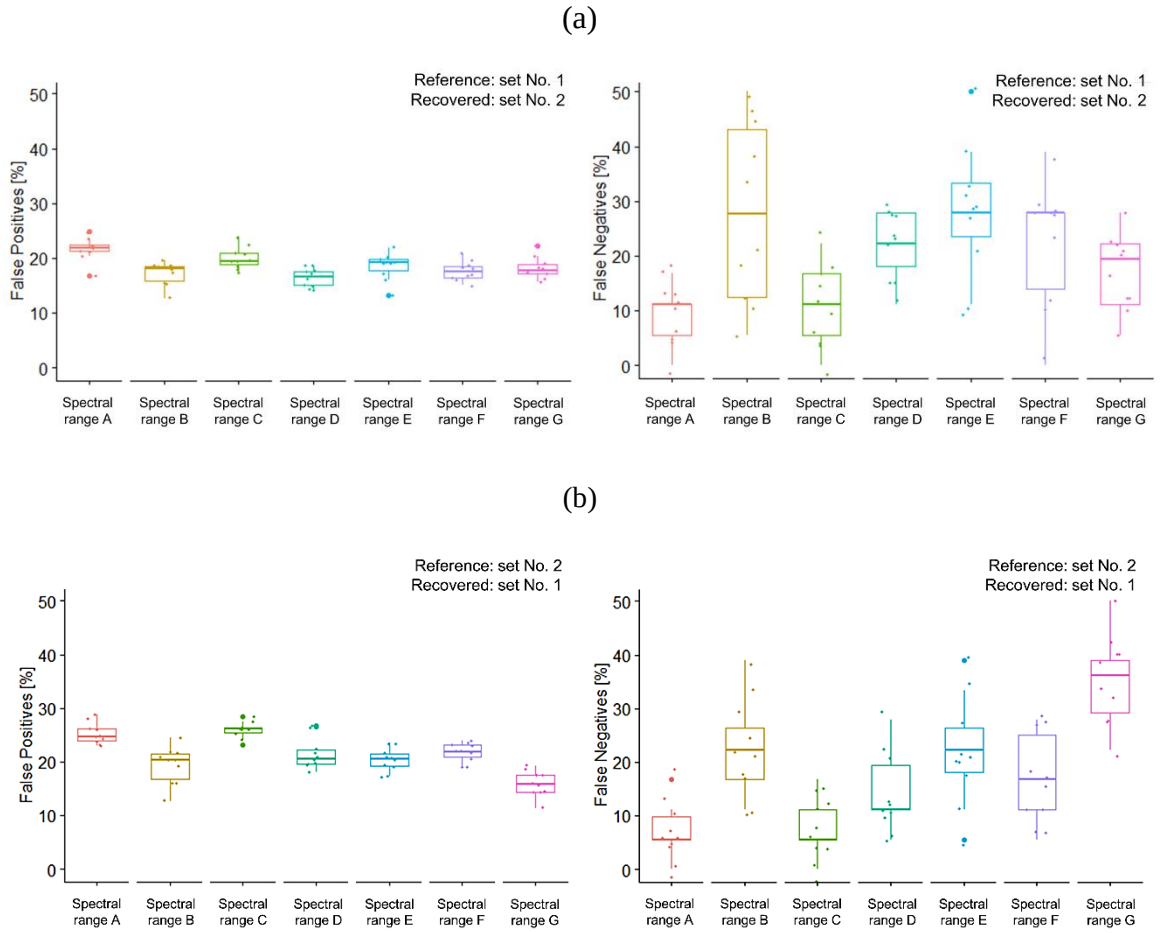


Fig. 4. The levels of false positive and false negative responses of LR models constructed for Raman spectra of bloodstains prepared using set No. 1 (a) and set No. 2 (b), respectively, for training the LR models.

The selection of preprocessing approach was individually determined by the type of distortion present in the recorded signals. As shown in Table 2, GA selected a different combination of preprocessing methods for almost each developed LR model, ensuring their robustness. One can also notice that denoising and baseline correction methods practically did not indicate the superiority of any specific methods with respect to others. Just a slight trend is observed among the normalization methods. The false negative responses (Fig. 4) observed for individual LR models revealed that the lowest error rates were associated with models developed on the following spectral ranges: 1200–1410 cm^{-1} and 1200–1300 cm^{-1} combined with 950–1020 cm^{-1}. This observation is also supported by the empirical cross entropy (ECE) plots in Fig. 5 (ECE plots corresponding to remaining LR models are depicted in Fig. S2 and S3). LR models based on these two spectral ranges yield incomparably more satisfying ECE outcomes. In both cases (denoted as spectral ranges A and C in Table 2), PQN – not SNV – was used as the normalization method. The most logical conclusion would attribute this model performance difference to the varying spectral ranges used to construct the LR models. In other words, variants A and C, covering the ranges 1200–1410 cm^{-1} and 1200–1300 cm^{-1} combined with 950–1020 cm^{-1}, respectively, are characterized by bands that exhibit the most pronounced time-related variations (high

values of b^2/w^2 ratios). Changes in spectral characteristics, reflected mainly in such spectral features as hemoglobin aggregation (1255 cm^{-1} and 976 cm^{-1}) and oxidation markers (between 1300–1400 cm^{-1}), are well-recognized effects of the formation of Hb degradation products [8–13]. This also confirms the results from our previous study [13], where it was found that the aging of blood traces boils down primarily to the conversion of oxyHb to metHb and HC, followed by the aggregation of heme species.

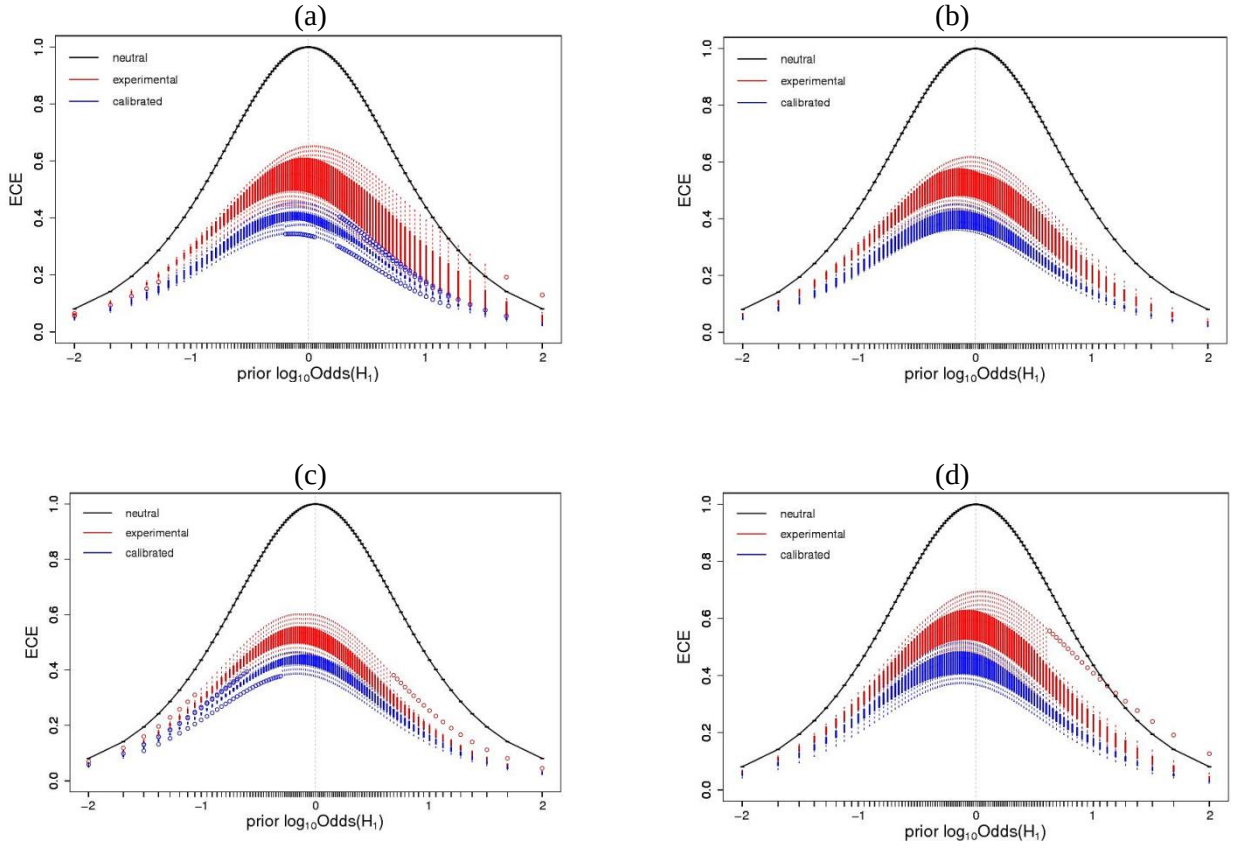


Fig. 5. Box-empirical cross entropy plots for the LR models constructed for Raman spectra of bloodstains prepared using the preprocessing strategies indicated by the genetic algorithm (GA). Two depicted LR models were trained on set No. 1 and tested using set No. 2 for the following spectral ranges: 1200–1410 cm^{-1} (a); 1200–1300 cm^{-1}, and 950–1020 cm^{-1} (b). The two remaining LR models were trained on set No. 2 and tested using set No. 1 for the following spectral ranges: 1200–1410 cm^{-1} (c); 1200–1300 cm^{-1}, and 950–1020 cm^{-1} (d).

3.1.2 Performance of the likelihood ratio models

Figure 4 and Table S2 portray the overall performance of the LR models under investigation, computed using the first (set No. 1) and the second (set No. 2) databases. The general findings expose that false positive (F_P) rates obtained for each of the seven model variants are somewhat more comparable to each other than the levels of false negative (F_N) answers. F_P oscillated around 20%, and the dispersion of results within each model is not as evident as in F_N.

This observation can be explained by the higher similarity of compared bloodstain samples at more advanced stages of degradation. Inconsiderable differences between the spectral characteristics of blood traces, a consequence of a slow-down in the aging process, appear to equally affect the LR models'

capability to distinguish between differently-aged samples, leading to the increase in the levels of F_P answers. The observed misleading support for the hypothesis about the common source of bloodstains (H_1: the recovered and reference bloodstains are characterized by the same TSD) obviously cannot be reduced by implementing other preprocessing techniques or signal manipulation in general. This is because the factor determining the F_P answers here is the rate of blood degradation itself and the ability to monitor it with a given analytical technique. Therefore, it needs to be clearly stated that the purpose of the present research was not to demonstrate the superiority of Raman spectroscopy over any other analytical method capable of characterizing the degradation state of bloodstains. Its objective was solely to evaluate the performance of the novel dating framework, which – eventually – proved effective when combined with Raman spectroscopy, especially at the initial stage of aging (as depicted in Fig. 6).

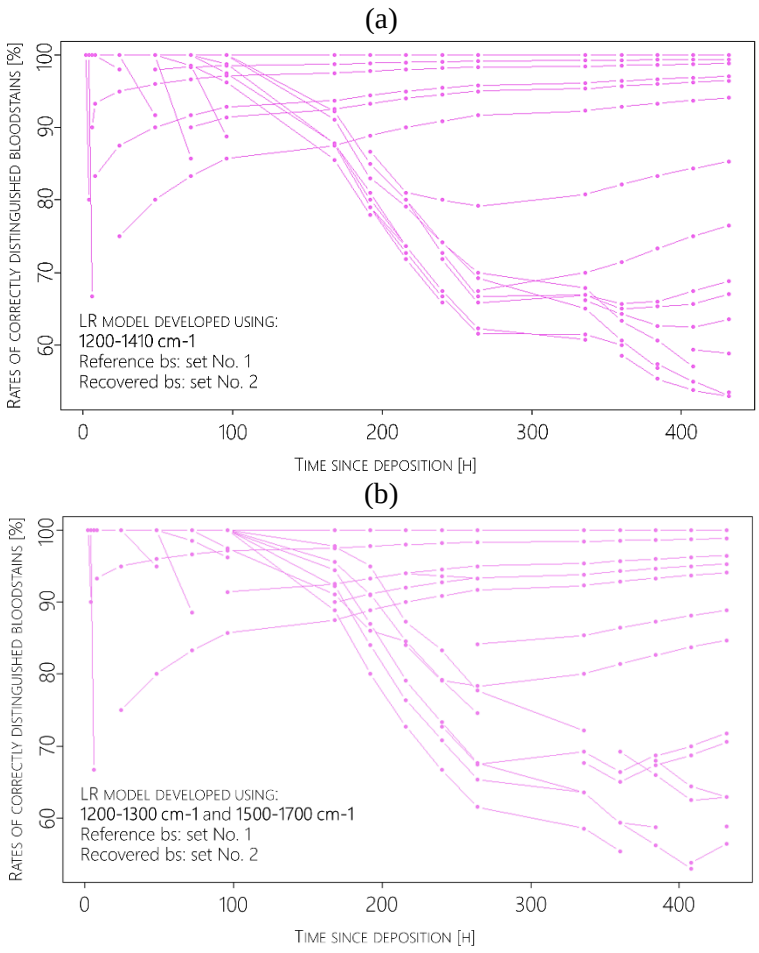


Fig. 6. The percentage of correctly distinguished differently-aged blood traces in the function of aging time [h] obtained when using LR models developed for spectral ranges A (a) and B (b). In both cases, the reference and recovered bloodstains originated from sets No. 1 and No. 2, respectively.

Reporting F_P responses over the entire analyzed aging period may be misleading due to its low information content. It should be clear that the longer the degradation time, the higher similarity of spectroscopic signals; hence the higher percentage of incorrect answers will be obtained. Consequently, it is much more reasonable to present the F_P rates concerning the time elapsed since bloodstains

deposition, as it would better report on the effectiveness of the dating approach. An example of such dependency for LR models developed for spectral ranges A and B is shown in Figures 6a and 6b, respectively. In both presented cases, LR models were trained using set No. 1. Instead of F_P responses, the plots depict the percentage of correctly distinguished differently-aged blood traces plotted against the aging time. Thus, one needs to subtract the value shown in the plot for a given time point from 100% to obtain the percentage of false positives. For example, for the first time point considered (two hours elapsed since bloodstain formation), the average percentage of correct discrimination is 98.82%. Thus, the corresponding F_P value corresponds to 1.18%. What is more, the primary purpose of Figure 6 is to demonstrate some time trends characterizing the performance of developed LR models. This time trend indicates nearly 100% discrimination effectiveness of the LR models within the first seven days of degradation, decreasing after about 170 hours, which is a benchmark performance. As can be seen, even though the overall performance of these models is entirely different (due to F_N rates, 9.45% and 27.77% for models A and B, respectively), their ability to discriminate between differently-aged samples over time correctly – reflected in the rates of F_P responses – is exceedingly high and relatively similar.

Consequently, the factor that actually differentiates the developed LR models concerning their effectiveness in solving the considered comparison problem is the already mentioned reduction of within-source variations, reflected in the percentage of false negative answers. After inspection of F_N rates, it becomes clear that LR models based on PQN-normalized spectral ranges A (1200–1410 cm^{-1}) and C (1200–1300 cm^{-1} combined with 950–1020 cm^{-1}) outperform the remaining LR models. F_N rates for these two models' variants (bolded in Table S2) are less dispersed and oscillate around 10% while they reach ca. 30% for the remaining spectral ranges. Thus, once again, it turns out that from the dating perspective, the most informative Raman features are the Hb aggregation markers (e.g., 1255 cm^{-1} or 976 cm^{-1}) and so-called oxidation markers located between 1300–1400 cm^{-1} [48–50]. Interestingly, the greater the influence of the 1500–1700 cm^{-1} range in the developed model, the worse its performance. An example of this detrimental effect of including the 1500–1700 cm^{-1} range in the model is the one based on spectral range B, namely 1200–1300 cm^{-1} combined with 1500–1700 cm^{-1}. This may suggest that bands appearing in this range, which contains the core-size and spin state markers [13], are more susceptible to random variation (e.g., resulting from dietary factors) than changes due to bloodstains aging. Hence, future studies should primarily focus on identifying factors influencing the degradation kinetics to simplify the process of supervised aging.

Finally, the ECE plots confirmed the conclusions reached upon examination of the levels of F_N answers, proving that LR models based on spectral ranges A (1200–1410 cm^{-1}) and C (1200–1300 cm^{-1} combined with 950–1020 cm^{-1}) produce the most desirable performances. The ECE plots corresponding to those two LR models are presented in Fig. 5. The plots for the five remaining spectral ranges (B and D–G) are provided in SM (Fig. S2 and S3). The experimental and calibrated curves (indicated in red and blue, respectively) are represented by boxplots positioned at the considered prior odds (prior probability quotients $Pr(H_1)/Pr(H_2)$). Each boxplot accounts for the ECE values calculated for a given

prior odds using the available likelihood ratio values. For the best-performing models, the reduction of information loss (or gain of information) reaches even ca. 50% (e.g., Fig. 5c), which is more than an acceptable outcome given the small databases. In other words, it means that after the bloodstain evidence analysis using developed LR models, a relatively large amount of information concerning the uncertainty about the correct hypotheses was explained. The ECE values also evidenced that the models based on 1500–1700 cm^{-1}-containing regions (Fig. S2 and S3) represent the worst solutions indisputably. Not only did they deliver the highest rates of false negative answers, but they also yielded incomparably inferior ECE outcomes.

It should also be noted that in the case of poorly performing models, the ECE plots deviated from the desired bell-shaped curves. The ECE values went beyond the neutral curve, especially for the positive logarithm of the prior odds, i.e., $\log_{10}\text{Odds}(H_1) > 0$. This deterioration may be caused by only a single sample delivering strong misleading support for the incorrect hypothesis (in this case – H_1). In particular, when small databases are considered, the penalty assigned to a misleading LR value might outweigh the importance given to the LRs supporting the correct hypothesis. For this reason, these inferior ECE results are not necessarily evidence of poorly performing LR models. The ECE approach may simply not be the most robust performance measure. Thus, it should also be interesting to explore that issue further in the future, trying to verify whether the obtained poor LR model performance, assessed using the ECE approach, may be caused by just a single misleading value. Moreover, some differences between the experimental and calibrated ECE plots obtained even for well-performing models (Fig. 5) hold promise for improving the models' effectiveness. The simplest way to refine the model may be to expand its databases to capture better all the relevant features characteristic of the whole population of the analyzed bloodstains.

4. Conclusions and future perspectives

The objective of the presented research was to verify the possibility of estimating the age of blood deposits – characterized by Raman signatures – through their comparison with "artificially" created reference materials. Thus, the conventional quest for the bloodstains dating method, based on regression analysis, was replaced with a discrimination problem, solvable through the likelihood ratio approach.

Given the inherent variability of analyzed materials and the multidimensionality of the considered issue, the capability of established models to discriminate between differently-aged samples was the benchmark performance, providing preliminary evidence of the effectiveness of the proposed dating framework. The best models, founded on 1200–1410 cm^{-1} and 1200–1300 cm^{-1} combined with 950–1020 cm^{-1} ranges, correctly distinguished nearly 100% of differently-aged blood traces within the first week of degradation. Additionally, across the monitored time range, these models delivered approximately 10% of false negatives and 20% false positive answers, while the ECE plots evidenced ca. 50% reduction in information loss. The false negative rates were gratifying, while the decent ECE

results, at this point, should not be too much of a concern. As already mentioned, when combined with scarce databases, ECE often leads to misleading conclusions due to its sensitivity to individual LR values supporting the incorrect hypothesis. In contrast, the 20% false positives may not be entirely satisfactory. Unfortunately, the level of false positive answers is not the product of improper preprocessing or data modeling but rather an effect of a significant slowdown in bloodstain aging processes that could no longer be effectively monitored by Raman spectroscopy. Hence, it should be concluded that Raman spectroscopy cannot be considered a panacea for tracing blood degradation processes and, in fact, was never considered one. It might be unrealistic to expect a single method to meet analytical requirements posed by any possible evidential bloodstain. In this study, Raman spectroscopy was only used to verify the effectiveness of LR models designed to discriminate differently-aged blood traces. This effectiveness – still very high during the first week of degradation (nearly 100% correct discriminations) – will increase once other analytical methods are applied, better suited to tracking aging changes in bloodstains. Thus, an interesting way forward may be inviting other researchers for the joint development of this approach by implementing other analytical techniques capable of probing the state of bloodstains degradation over extended periods (e.g., FT-IR, UV-Vis spectroscopies, or even RNA analysis), which is now the subject of our further research.

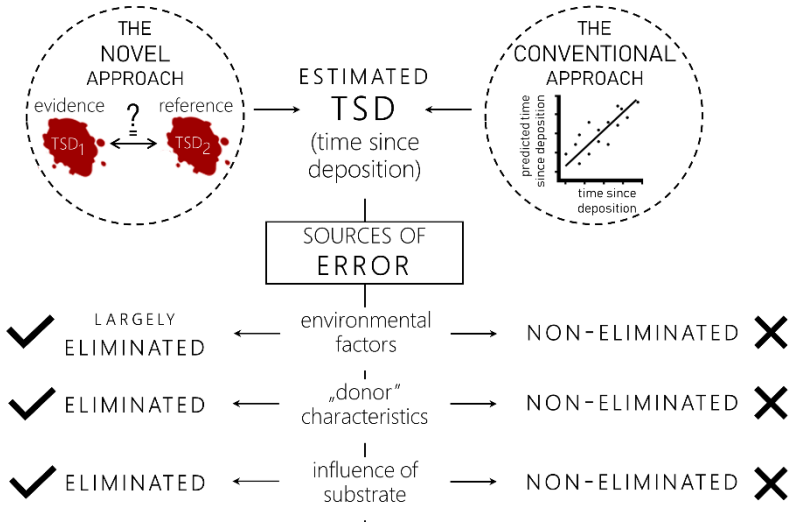


Fig. 7. The novel approach *versus* the conventional dating approach – sources of error.

Another important note for future studies is that the high reproducibility of the supervised aging procedure is the determining factor for reliable and valid dating. In other words, a similar volume of blood, freshly drawn from the donor (if possible, the same individual who was the donor of the questioned trace), should be deposited on an identical substrate and stored under similar conditions as those observed at the crime scene. To reach valid TSD estimates, bloodstains serving as reference samples for building models must replicate the aging process of the evidence as accurately as possible. Irrespective of the monitoring method applied, the bloodstains' resultant aging profile depends on the aging kinetics, which is affected by such influential factors as the substrate, environmental conditions

persisting during aging (controlled by climatic chambers), and the initial composition of the bloodstain. Therefore, the success of the proposed approach will depend on how much the above factors influence degradation and, above all, on how much it is possible to control them during supervised aging. This way, sources of error in TSD estimates that occur during the conventional dating approach would be largely eliminated (Fig. 7).

Indeed, the mere statement that errors arising from external factors will be largely eliminated is not a sufficient guarantee of the reliability of the proposed methodology. Hence, once again, it should be reminded that to meet the ISO/IEC 17025 international quality standards [16] or – simply – to ensure the validity of any novel methodology [51–53], a forensic laboratory must undertake a comprehensive evaluation of all factors that could impact the examination result during the validation process of a method intended for use in casework. In this specific case, after identifying the factors (and their combinations) that influence the degradation of bloodstains, the accuracy of TSD estimates using the new methodology can be fully assessed. It will be achieved by validating likelihood ratio models, including determining the rates of false positive and false negative responses, for "evidential" bloodstains degrading under different conditions, not only those presented in the present study. Only then will it be possible to consider advancing this methodology to practical application.

Obviously, it would be somewhat naïve to assume the possibility of recreating precisely the same conditions as evidence degradation. Nonetheless, recreating them as closely as possible to those prevailing during the degradation of the evidential bloodstain may be sufficient to answer the prosecutor's questions regarding the time of traces' formation. Thus, in the first place, future research should focus on identifying factors that may affect the bloodstains' degradation – using experimental design methods [54] – so that the supervised aging procedures focus on controlling only those specific factors. One of the most significant and likely most challenging tasks will be understanding the relationship between the initial blood composition (the "donor" characteristics) and the kinetics of its degradation. Determining this relationship will allow answering whether the supervised aging procedure will need to be carried out using blood from the person of interest/suspect or whether a comparison sample could be constructed from any blood. In the latter case, the proposed dating procedure would still be feasible, especially if the suspect was actually guilty of the alleged crime and refused to provide a blood sample for testing. In a much more likely scenario, when the source of blood will determine the rate of aging, it is worth remembering that forensic science can be used not only to prove the guilt but also to exonerate innocent people from the charges brought against them. In the era of social doubts about the reliability of forensic sciences, it is important to avoid jailing people because of flawed or insufficient forensic techniques. The inability to estimate the time of bloodstain formation creates an analytical gap and carries the risk of convicting an innocent person. In other words, in cases where a blood sample is to be obtained from the wrongly accused person and could ultimately lead to one's acquittal, the problem of obtaining blood for comparative analysis should disappear.

Finally, it is important expressly to highlight that the case-by-case approach founded on the idea of comparing the evidence with some reference materials should be of a wider forensic application. The proposed framework is expected to enrich the forensic arsenal of analytical tools when answering the age of other evidence, such as different types of body fluids, organic gunshot residues [55], or fingerprints [56]. In this regard, this novel dating approach is quite universal. It just requires to be complemented with an analytical technique adequate for characterizing aging profiles of the forensic trace of interest, while the fundamental idea of dating – founded on the LR-based comparison – remains the same.

Funding

The research did not recive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

There are no conflicts to declare.

References

1 C. Weyermann, O. Ribaux, *Sci. Justice*, 2012, **52**, 68–75.
2 O. Ribaux, S. Walsh, P. Margot, *Forensic Sci. Int.*, 2006, **156**, 171–181.
3 O. Ribaux, A. Baylon, C. Roux, O. Delémont, E. Lock, C. Zingg, P. Margot *Forensic Sci. Int.*, 2010, **195**, 10–16.
4 L. Tomellini, *Arch. d'Antropologie criminelle de Criminol.*, 1907, **14**, 2.
5 R.H. Bremmer, K.G. de Bruin, M.J.C. van Gemert, T.G. van Leeuwen, M.C.G. Aalders, *Forensic Sci. Int.*, 2012, **216**, 1–11.
6 G. Zadora, A. Menżyk, *TrAC*, 2018, **105**, 137–165.
7 V. Sharma, R. Kumar, *TrAC*, 2018, **107**, 181–195.
8 K. M. Marzec, A. Rygula, B. R. Wood, S. Chlopicki, M. Baranska, *J. Raman Spectrosc.*, 2014, **46**, 76–83.
9 P. Lemler, W.R. Premasiri, A. DelMonaco, L.D. Ziegler, *Anal. Bioanal. Chem.*, 2014, **406**, 193–200.
10 K.C. Doty, G. McLaughlin, I.K. Lednev, *Anal. Bioanal. Chem.*, 2016, **408**, 3993–4001.
11 K. C. Doty, C. K. Muro, I. K. Lednev, *Forensic Chem.*, 2017, **5**, 1–7.
12 H.J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N.J. Fullwood, B. Gardner, P.L. Martin-Hirsch, M.J. Walsh, M.R. McAinsh, N. Stone, F.L. Martin, *Nat. Protoc.*, 2016, **11**, 664–687.
13 A. Menżyk, A. Damin, A. Martyna, E. Alladio, M. Vincenti, G. Martra, G. Zadora, *Talanta*, 2020, **209**, 120565.
14 H. Lin, Y. Zhang, Q. Wang, B. Li, P. Huang, Z. Wang, *Sci Rep.*, 2017, **7**, 13254.
15 R. Kumar, K. Sharma, V. Sharma, *Sci. Justice*, 2020, **60**, 538–546.
16 C. Gannicliffe, End User Commentary on Investigating the Age of Blood Traces: How Close Are We to Finding the Holy Grail of Forensic Science?, in: S. Francese (Ed.), Emerging Technologies for the Analysis of Forensic Traces, *Springer*, 2019, pp. 129–132.

17 G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, *Wiley*, Chichester, 2014.
18 ENFSI Guideline for Evaluative Reporting in Forensic Science. Strengthening the Evaluation of Forensic Results across Europe. European Network of Forensic Science Institutes, 2015, https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf, (accessed 31st August 2021).
19 A. Martyna, G. Zadora, Hybrid Likelihood Ratio Models for Forensic Applications: a Novel Solution to Determine the Evidential Value of Physicochemical Data, in: L. Dennany (Ed.), Challenges in Detection Approaches for Forensic Science, *RSC*, 2021, pp. 198–231.
20 A. Damin, et al., European patent No. WO2017077513 (A1), (2017).
21 M. Signorile, F. Bonino, A. Damin, S. Bordiga, *Top. Catal.*, 2018, **61**, 1491–1498.
22 A. Martyna, G. Zadora, T. Neocleous, A. Michalska, N. Dean, *Anal.Chim. Acta*, 2016, **931**, 34–46.
23 A. Martyna, K.-E. Sjastad, G. Zadora, D. Ramos, *Talanta*, 2013, **105**, 158–166.
24 A. Martyna, A. Michalska, G. Zadora, *Anal. Bioanal. Chem.*, 2015, **407**, 3357–3376.
25 A. Michalska, A. Martyna, J. Zieba-Palus, G. Zadora, *J. Raman Spectrosc.*, 2015, **46**, 772–783.
26 A. Martyna, G. Zadora, D. Ramos, *J. Anal. Appl. Pyr.*, 2018, **133**, 198–215.
27 J. Engel, L. Blanchet, B. Bloemen, L. Heuvel, U. Engelke, R. Wevers, L. Buydens, *Anal. Chim. Acta*, 2015, **89**, 1–12.
28 A. Martyna, A. Menżyk, A. Damin, A. Michalska, G. Martra, E. Alladio, G. Zadora, *Chemom. Intell. Lab. Syst.* 2020, **202**, 104029.
29 O. Kvalheim, F. Brakstad, Y.-Z. Liang, *Anal. Chem.*, 1994, **66**, 43–51.
30 S. Mallat, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1989, **7**, 674–693.
31 A. Savitzky, M. Golay, *Anal. Chem.*, 1964, **36**, 1627–1639.
32 P. Eilers, *Anal. Chem.*, 2003, **75**, 3631–3636.
33 Z. Zhang, S. Chen, Y. Liang, *Analyst*, 2010, **135**, 1138–1146.
34 . Zhang, S. Chen, Y. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, H. Zhou, *J. Raman Spectrosc.*, 2010, **41**, 659–669.
35 P. Cadusch, M. Hlaing, S. Wade, S. McArthur, P. Stoddart, *J. Raman Spectrosc.*, 2013, **44**, 1587–1595.
36 J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan, *Anal. Chim. Acta*, 2010, **683**, 63–68.
37 A. Ruckstuhl, M. Jacobson, R. Field, J. Dodd, *J. Quant. Spectrosc. Radiat. Transf.*, 2001, **68**, 179–193.
38 C. Ryan, E. Clayton, W. Griffin, S. Sie, D. Cousens, *Nucl. Instrum. Methods Phys. Res. B*, 1988, **34**, 396–402.
39 C. Lieber, A. Mahadevan-Jansen, *Appl. Spectrosc.*, 2003, **57**, 1363–1367.
40 J. Zhao, H. Lui, D. McLean, H. Zeng, *Appl. Spectrosc.*, 2007, **61**, 1225–1232.
41 L. Komsta, *Anal. Bioanal. Chem.*, 2014, **406**, 1985–1998.
42 R. J. Barnes, M. S. Dhanoa, S. J. Lister, *Appl. Spectrosc.*, 1989, **43**, 772–777.
43 F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, *Anal. Chem.*, 2006, **78**, 281–4290.
44 B. W. Silverman. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, UK, 1986.
45 D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, *J. Forensic Sci.*, 2013, **58**, 1503–1518.
46 R. Haraksim, D. Ramos, D. Meuwly and C. E. H. Berger, *Forensic Sci. Int.*, 2015, **249**, 123–132.
47 R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018. URL, https://www.R-project.org/.
48 B. R. Wood, L. Hammer, D. McNaughton, *Vib. Spectrosc.*, 2005, **38**, 71–78.
49 B. R. Wood, L. Hammer, L. Davis, D. McNaughton, *J. Biomed. Opt.*, 2005, **10**, 14005–14013.

50 B. R. Wood, P. Caspers, G. J. Puppels, S. Pandiancherri, D. McNaughton, *Anal. Bioanal. Chem.*, 2007, **387**, 1691–1703.
51 A. Sikirzhytskaya, V. Sikirzhytski, L. Perez-Almodovar, I. K. Lednev, *Forensic Chem.*, 2023, **32**, 100468.
52 N. A. Nichols, I. K. Lednev, *Am. J. Anal. Chem.*, 2022, **13**, 1–8.
53 Y. V. Kistenev, A. V. Borisov, A. A. Samarinova, S. Colon-Rodriguez, I. K. Lednev, *Sci. Rep.*, 2023, **13**, 5384.
54 R. Leardi, *Anal. Chim. Acta*, 2009, **652**, 161–172.
55 S. Charles, N. Geusens, E. Vergalito, B. Nys, *Forensic Sci. Int.: Synergy*, 2020, **2**, 416–428.
56 K. Wertheim, *J. Forensic Identif.*, 2003, **53**, 42–49.

Breaking with trends in forensic dating: A likelihood ratio-based comparison approach

Supplementary Material available.

Abstract

Further steps toward understanding the time-related information contained within bloodstains found at the crime scene are rightly considered a top priority in forensic science. Contrary to widely held assumptions, the reason for the delayed exploitation of bloodstains dating methods in practice is not the lack of suitable analytical techniques for monitoring degradation processes. The problem lies in the variability of the environmental and circumstantial conditions, playing a vital role in the degradation kinetics of blood deposits. The present article demonstrates the possibility of breaking with current approaches based on absolute age estimations to finally answer time-centered questions in real forensic scenarios. The proposed novel framework for situating forensic traces in time is based on the likelihood ratio assessment of the (dis)similarity between the evidence decomposition and sets of reference materials obtained through supervised aging. In such a strategy, every dating procedure is constructed on a case-by-case basis to fit examined blood traces, thereby limiting the adverse influence of external factors on the validity of age estimations and providing a way for future crime scene implementation.

Keywords: Bloodstains; Time since deposition, Forensic dating, Likelihood ratio, Comparison problem

1. Introduction

Forensic experts – nowadays equipped with DNA profiling methods – have never been better prepared to answer questions about a suspect's identity, which usually arise during the investigation process. Genetic typing of biological evidence – such as bloodstains – has clearly revolutionized forensic science to such an extent that it is now one of the cornerstones of modern policing. However, in the era of DNA testing, it may be overlooked that identifying the donor of the collected biological trace is not always the most critical issue. Thus, to demonstrate a possibly strong link between the evidence and the investigated crime, it is often necessary to prove that the questioned material was created during the incriminating event [1]. This can be achieved by providing information about the time elapsed since trace deposition (TSD, time since deposition).

A reliable answer to the question of bloodstains' age could be a significant added value to the criminal inquiry, allowing the judicature to recreate a criminal situation *a posteriori* [2, 3]. Information about the time of bloodstain(s) formation may prove helpful directly during the investigative phase of the criminal procedure, for example, by establishing a timeline of events through sequencing the deposition of traces, but not only. The time-centered questions often call for rational responses also at the evaluative stage of the proceeding when hypotheses representing the prosecutor's and defense's standpoints are confronted. At this phase, a common defense tactic is to question the validity of evidential bloodstains deposited by the suspect at the crime scene. In such a case, the prosecutor's opponents do not challenge the identification of blood traces but rather the time of their formation, thereby casting doubts on the suspect's presence at the crime scene during the incriminating event. In this context, the role of forensic experts is to indicate which hypotheses presented by both parties to the proceedings are supported by the evidence, namely the age of the questioned blood traces. Consequently, correct estimation of TSD could help verify bloodstains' relevance – and hence also the suspect's connection – to the case considered in the courtroom scenario.

Awareness of the gravity of this temporal information – or lack thereof – has accompanied forensic experts and judiciary members for at least a hundred years [4]. Nevertheless, despite significant research efforts [5–7], a reliable method for bloodstains dating is still unavailable. The root cause of the problem is not the researchers' inability to develop an analytical tool capable of monitoring the time-dependent properties of bloodstains, as evidenced by a constantly growing body of research in forensic dating [5–7]. Indeed, spectroscopic methods, particularly Raman spectroscopy, have contributed to a better understanding of degradation mechanisms inside blood traces [6, 8–15]. However, despite these remarkable technological advancements, none of the dating approaches has yet to overcome the experimental research phase. No forensic organization working according to ISO/IEC 17025 international quality standards performs the absolute dating of bloodstains on a daily analytical basis [16]. This proves that obtaining sound analytical results does not accurately address the entire problem,

as the mere chemical information about the aging process of blood – no matter how detailed – turns out to be insufficient if not integrated into an appropriate dating framework.

A closer inspection of previous research suggests that the delayed exploitation of the developed dating models in forensic practice might depend on the data analysis approach. According to the commonly adopted strategy (hereinafter referred to as the *conventional dating approach*), most proposed methods have sought the dependency between the TSD and some dynamic properties of blood, degrading under stable, strictly controlled conditions through (multivariate) regression analysis [5, 6]. However, apart from yielding dating models of excellent performance in laboratory settings, the conservative practice of controlling factors affecting the aging kinetics (e.g., environmental conditions) may pose certain risks. These models – consistently fitting the aging behavior of training samples used for their construction – are not equally effective in the typical multivariate and hardly predictable forensic practice. In other words, forensic experts are faced here with the age-old question of the transferability of proposed conventional dating methods, undoubtedly powerful in artificial laboratory settings, to more realistic contexts.

External factors will always affect the blood degradation process, altering its rate to a lesser or greater extent. This means that when dealing with evidence, the uncertainty deriving from these influencing agents cannot be eliminated or universally estimated in advance. It should be incorporated into the dating strategy. Such a shift in perspective is proposed in the present work. According to the strategy adopted in this study, verifying the relevance of blood evidence to the considered crime should not necessarily be performed using previously established calibration models. A strong dependency between the accuracy of TSD estimations and the aging kinetics of blood at the crime scene constitutes a major contraindication to adopting this conventional approach. It is hypothesized instead that in order to situate the trace in time, it suffices to compare the evidential material with reference bloodstains, created in such a way as to match the local conditions and time since deposition (TSD) corresponding to the prosecution's/defense's version of events (e.g., t_R in Fig. 1).

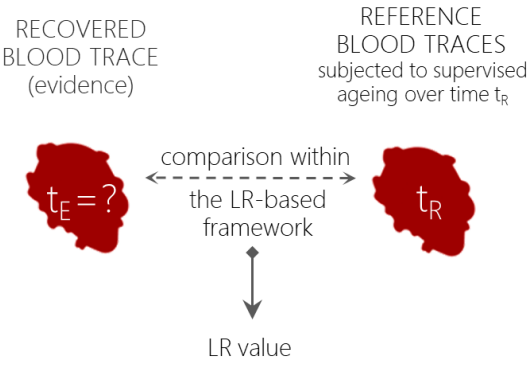


Fig. 1 Implementation of the comparison procedure within the LR framework to estimate the TSD of a questioned bloodstain.

The critical stage of this comparison is the likelihood ratio-based (LR-based) assessment [17–19] of the (dis)similarity between the analytical signals reflecting the decomposition stage of evidence

and reference materials obtained through supervised aging. This process is designed to recreate the degradation of the evidence on site as precisely as possible to normalize the aging kinetics of compared materials (evidence and reference bloodstains). It means that every dating procedure would be constructed case-by-case, each time tailored to fit the examined traces. In such a way, the influence of external factors (e.g., environmental conditions) on the aging kinetics and, through that, on the analysis's validity should be considerably reduced, providing a way for future crime scene implementation.

2. Materials and methods

2.1 Bloodstains databases

Since the main idea behind the novel dating framework developed herein is the LR-based comparison of recovered (evidence) and reference bloodstains (samples of known age), at least two databases were required for the practical proof of concept. One set was the source of data for training the LR model and deriving the reference sample of bloodstain (the product of the so-called supervised aging process) and the second set from which the recovered bloodstain (acting as the evidence) was selected. Eventually, two separate databases (hereafter referred to as sets No. 1 and 2) were prepared at a three-weeks time distance, according to the scheme presented in Fig. 2.

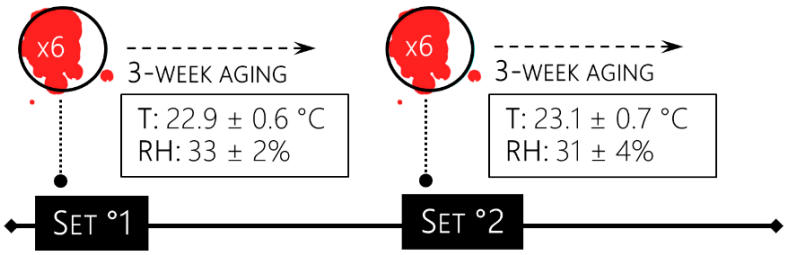


Fig. 2 The general scheme of the supervised aging process of two different bloodstain sets served to develop and validate LR models.

Each database consisted of spectral signatures of degrading blood traces created by depositing 20 µL of additive-free capillary blood (derived from a single donor) in aluminum sample pans. During the first day of degradation, blood deposits (six per database) were analyzed for up to eight hours elapsed since bloodstain formation with a two-hour interval. Due to the significant slow-down in the degradation process [13], the time resolution of measurements was subsequently reduced, and spectra were registered daily (excluding weekends) for the next three weeks. This measurement protocol resulted in databases consisting of spectra corresponding to 18 time-points, assumed to constitute 18 different evidence time-related sources (*source* should be understood as a group of bloodstains corresponding to the same TSD). Each time-point was characterized by six bloodstains (which served as replicate measurements of samples at a given time-point) created one after another at appropriate time intervals, estimated with regard to the duration of spectral acquisition. In order to meet the criteria of supervised aging, between measurements, all samples were stored in darkness, exposed to temperature (T) and

relative humidity (RH) controlled by the laboratory air-conditioning system to normalize the degradation kinetics between each dataset as much as possible (Fig. 2).

It should also be added that within this work scope, Raman spectroscopy (RS) was selected to characterize the state of bloodstains degradation. The rationale for this choice was given in our previous work [13] that demonstrated the capability of RS to deliver information inherent to chemical changes accompanying the *ex vivo* aging of blood. Raman signatures were obtained using a Renishaw *inVia Raman Microscope,* following the procedure developed in [13]. To prevent sample damage from the excessive point laser irradiation and increase the representativeness of a single measurement, all spectra were registered during the magnetic-driven rotation of the sample [20, 21]. Detailed experimental conditions used are summarized in Table 1.

Table 1 The rotating mode experimental conditions used during the spectral characterization of bloodstains.

Experimental conditions	
Laser excitation source [nm]	785
Laser density [mW/μm^2]	ca. 0.16 (10% of total laser power)
Spectral range [cm^{-1}]	600–1800
Objective	5× NIR optimized objective (N.A. = 0.1)
Accumulations	2
Time of acquisition [s]	20

2.2 The likelihood ratio procedure for bloodstains comparison

In judicial practice, verifying the association between the evidence and reference materials to establish whether they could originate from the same (H_1) or two different sources (H_2) is defined as a classical comparison problem [17–19]. In this study, however, we are dealing with its unusual variation (depicted in Fig. 1), where the *source* should be understood as a group of bloodstains corresponding to the same TSD. Thus, the considered hypotheses can be formulated as follows:

H_1: the age of the recovered evidence (questioned bloodstain(s) revealed at the crime scene; t_E) MATCHES the age of the reference material (bloodstains obtained during the process of supervised aging according to the prosecutor's or defense standpoint; t_R): $t_E = t_R$.

H_2: the age of the recovered evidence (questioned bloodstain(s) revealed at the crime scene; t_E) DOES NOT MATCH the age of the reference material (bloodstains obtained during the process of supervised aging according to the prosecutor's or defense standpoint; t_R): $t_E \neq t_R$.

Even though such a comparison implies a discrimination task, easily solvable with chemometric tools, in actuality, establishing the (dis)similarity of bloodstains by comparing the likeness of the data solely (e.g., Raman spectra) is insufficient. It should be reminded that adequately performed forensic expertise also involves interpreting the findings to establish their evidential value. For the forensic assessment of evidence within a comparison problem, it is essential not only the similarity but also the

rarity of the compared patterns in the relevant population. Thus, emphasis has to be put on the question: what does observed similarity mean in this particular case?

To answer that question, the likelihood ratio (LR) framework [17–19] is recommended, which considers not only the similarity and the rarity of physicochemical characteristics but also the possible levels of variation (within- and between-source variability). The LR quantifies the strength of sources' similarity, which escalates with their increasing rarity, indicating how strongly they are alike to verify whether they share a common source. It is computed as the probability of observing the physicochemical data for the samples (E), given the propositions (H_1 and H_2):

$$\boldsymbol{LR} = \frac{\mathbf{Pr}(\boldsymbol{E}|\boldsymbol{H_1})}{\mathbf{Pr}(\boldsymbol{E}|\boldsymbol{H_2})}.$$

The LR can have a value between zero and infinity, and its interpretation is the following: LR values above one support H_1, while LR values below one provide support for H_2. Moreover, the higher (lower) the LR value, the stronger the H_1 (H_2) support. Therefore, LR is a qualitative and quantitative measure of the strength of the support for one of the hypotheses, which – in the forensic context – is a highly desirable feature.

The conventional LR approach, however, is not well-adjusted to deal with high-dimensionality data [17] where the number of variables (*J*) exceeds the sample population's size (*I*). When $J > I$, the LR models fail due to the singularity or the instability of the inverse of some variance-covariance matrices; thus, reducing data dimensionality becomes obligatory. A remedy for this obstacle is a form of symbiosis between the LR approach and chemometrics known as hybrid LR models, introduced by Martyna *et al.* [22], which have proven their effectiveness through years of use in forensic studies [23–26]. Reduction of data dimensionality and extraction of informative features (concerning the purpose of the study) are carried out through the implementation of chemometrics (*2.2.1 The chemometric tactic of data treatment*), resulting in the subsequent construction of LR models based on a limited number of new latent variables (*2.2.2 The development and validation of hybrid likelihood ratio models*).

2.2.1 The chemometric tactic of data treatment

Well-performing LR models for discrimination purposes are defined for just a few latent variables that maximize the between-source variation, $\boldsymbol{B}$, while minimizing the data variation within each source, $\boldsymbol{W}$. One method that allows finding these new latent variables is rMANOVA [27], whose task is to find the directions along which the between-source variance is the highest and the within-source variance is the lowest. These directions are defined as the eigenvectors of the matrix

$$[(1-\delta)W + \delta T]^{-1}B.$$

$\boldsymbol{T}$ is the target matrix, which is either $\boldsymbol{T} = \frac{\mathbf{1}}{\boldsymbol{p}}\boldsymbol{tr}(\boldsymbol{W})$ or $\boldsymbol{T} = \boldsymbol{diag}(\boldsymbol{W})$ when the variances of *p* variables for each source are equal or unique, respectively. In these studies, equal variances were assumed for each source to keep consistent with the LR model's assumption. The δ estimation depends

on the chosen target and evaluates the variance of the $\boldsymbol{W}$ components according to the Ledoit-Wolf theorem [27]. It should also become clear that the choice of rMANOVA was not accidental, as it demonstrates the possibility of solving two problems simultaneously. It reduces data dimensionality in a beneficial way for optimal source separation and highlights their specific features. For the above reasons, the LR models proposed for comparative dating of bloodstains were built for the first latent variable, LV1, from rMANOVA, which in our recent studies [27] proved reasonably effective in the separation of the sources described by highly multivariate data.

rMANOVA was performed for the signals subjected to adequate transformations to eliminate undesired spectral contributions – undermining variation of interest – and expose the vital features for further analysis. These mathematical transformations were effected through a suitable combination of preprocessing techniques applied to the raw signals, which – in the case of Raman spectra – usually includes denoising/smoothing, baseline correction, and normalization. In this research, the preprocessing aimed to support the LR models' performance by uniquely defining the characteristics of each source referred to different time points of bloodstains deposition. For this reason, a concept based on genetic algorithms (GA) previously developed by A. Martyna et al. [28] was implemented that remarkably supports the discrimination analysis of the signals owing to a case-shaped optimization of the preprocessing strategy. To measure the performance of the given combination of methods, reflected in the effectiveness of sources (differently-aged bloodstains) separation, a quality parameter (being also a fitness function) was defined as a ratio of the between-source and within-source variation (b^2/w^2) of the first latent variable (LV1) from the regularised MANOVA (rMANOVA) [27].

The applied preprocessing strategy was initialized with denoising/smoothing, followed by baseline correction and normalization. Additionally, to compensate for heteroscedastic noise [29] that grows with signal intensity and thus affects the covariance structure of the variables, denoised/smoothed signals, $\boldsymbol{a} = [\boldsymbol{a_1}, \ldots, \boldsymbol{a_J}]$ were subjected to the centered log-ratio transform (CLR) in the following manner:

$$\boldsymbol{s} = \boldsymbol{log_{10}a} - \boldsymbol{1/J}\sum_{\boldsymbol{i=1}}^{\boldsymbol{J}} \boldsymbol{log_{10}a_i}.$$

The most optimal preprocessing strategy was indicated using a GA-based procedure from among chromosomes, each described by three genes referring to denoising, baseline correction, and normalization steps. The tested denoising strategies involved 16 different methods based on discrete wavelet transform (DWT) [30] and Savitzky-Golay (SG) filter [31]. The baseline correction methods summed up to 64 various tools using asymmetric penalized least squares (AsPLS) [32–36], robust baseline estimation (RBE) [37], statistics-sensitive non-linear iterative peak-clipping (SNIP) [38], polynomials fitting [39, 40], and quantile regression (polyQR) [41]. Finally, standard normal variate (SNV) and probabilistic quotient normalization (PQN) [42, 43] were introduced among the available options of normalization methods. The groups of parameters, which characterize the considered preprocessing techniques, were roughly selected for optimization after a visual examination of the

preprocessed signals. Table S1 summarises all preprocessing methods implemented in the GA optimization procedure.

The initial generation in GA entailed 50 randomly selected preprocessing strategies. The chance of mutations was 0.1, the elitism level was 5%, and the algorithm converged when three subsequent solutions were identical.

2.2.2 The development and validation of hybrid likelihood ratio models

The objective of developed LR models is to compare the characteristics of the reference material from the first database (set No. 1) and the potential evidence – the recovered sample originating from set No. 2 – to assess whether they can be of the same (H_1) or different (H_2) age. As a reminder, considered hypotheses were defined in *2.2 The likelihood ratio procedure for bloodstains comparison*. In the beginning, the following procedure was proposed for comparing two bloodstains. Set No. 1 and 2 were subjected to the same preprocessing procedure, selected using the GA for the set, which collects the reference samples. Table 2 lists the selected preprocessing strategies, depending on whether the reference samples used for training the model were part of set No. 1 or 2. The first database (set No. 1), containing six Raman spectra recorded for 18 time-related bloodstains (6×18 signals), was divided into two parts. Three signals corresponding to 18 time-related points (3×18 signals) were randomly selected to constitute a sub-database (set No. 1a). The remaining created the second equally-sized sub-database (set No. 1b). Set No. 1a was used for summarising the case-specific data (providing information about bloodstains within each time-point) to find the new LV1 direction from rMANOVA, which best separates the spectra from different time-related points. The data from set No. 1b and the spectra measured for the recovered bloodstains (constituting set No. 2) were then mean-centered (using the mean of set No. 1a) and projected on LV1. Finally, the projections for the recovered sample, defined by the mean of $k_1 = 3$ data $\bar{\boldsymbol{y}}_{\boldsymbol{1}} = \sum_{i=1}^{k_1} \boldsymbol{y}_{1i}/\boldsymbol{k}_{\boldsymbol{1}}$ and one of the reference bloodstains, defined by a mean of $k_2 = 3$ data $\bar{\boldsymbol{y}}_{\boldsymbol{2}} = \sum_{i=1}^{k_2} \boldsymbol{y}_{2i}/\boldsymbol{k}_{\boldsymbol{2}}$ were compared. To define the relevant distributions, the remaining $M = 17$ samples with $n = 3$ data each, described by their means $\bar{\boldsymbol{x}}_{\boldsymbol{m}} = \sum_{j=1}^{n} \boldsymbol{x}_{mj}/\boldsymbol{n}$, were used as a background population.

The within-source variance was defined as:

$$\boldsymbol{w^2} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{n}(\boldsymbol{x}_{mj} - \bar{\boldsymbol{x}}_{\boldsymbol{m}})^2}{\boldsymbol{M(n-1)}},$$

while the between-source variance was expressed as:

$$\boldsymbol{b^2} = \frac{\sum_{m=1}^{M}(\bar{x}_m - \bar{x})^2}{\boldsymbol{M(n-1)}} - \frac{\boldsymbol{w^2}}{\boldsymbol{n}},$$

where $\bar{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{Mn}}\sum_{m=1}^{M}\sum_{j=1}^{n}\boldsymbol{x}_{mj}$.

The LR formula applied herein consists of four components, as shown in Fig. 3. Its first part (denoted in Fig. 3 as LR_1) expresses the probability density of the distribution with a mean equal to zero

and variance equal to $\frac{w^2}{k_1}+\frac{w^2}{k_2}$ at $\bar{y}_1-\bar{y}_2$. The second component (LR_2) estimates the probability density of the distribution estimated with Gaussian kernels positioned at each $\bar{x}_m$ with variance equal to $\frac{w^2}{k_1+k_2}+h^2b^2$ at $\bar{y}^*=\frac{\bar{y}_1k_1-\bar{y}_2k_2}{k_1+k_2}$. The h is expressed as $\left(\frac{4}{M(2p+1)}\right)^{\frac{1}{p+4}}$ where p is the number of the considered variables. In the case of this study, the p equals one. The final density curve is obtained by averaging Gaussian kernels to integrate into the unit area [44].

In contrast to the first two components, which assume that H_1 is true, the remaining two expressions in the denominator of the LR formula assume the opposite, so they are independent. They both estimate the probability density of the distribution estimated with Gaussian kernels positioned at each $\bar{x}_m$ with variance equal to $\frac{w^2}{k_1}+h^2b^2$ (or $\frac{w^2}{k_2}+h^2b^2$) at $\bar{y}_1$ (LR_3) or $\bar{y}_2$ (LR_4).

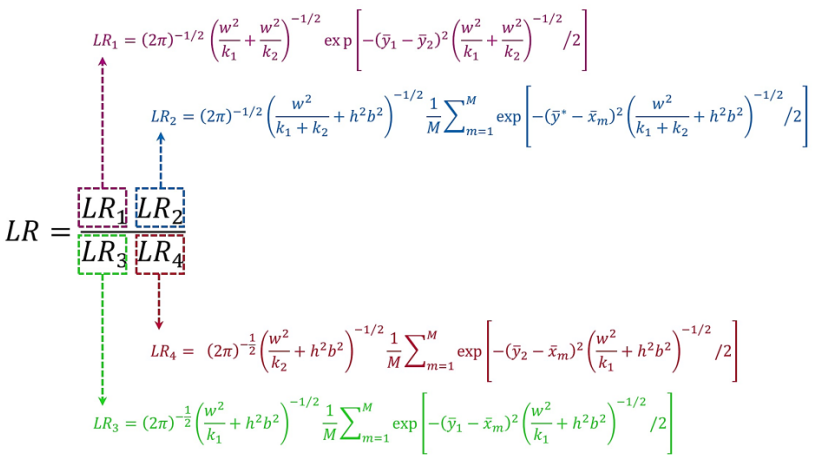


Fig. 3 The expression for the likelihood ratio (LR) applied in the study.

Subsequently, this division was reversed; the second database (set No. 2) was used to train the models, and the first database (set No. 1) provided the evidence samples for the comparison procedure. Models were constructed based on seven different ranges of the Raman spectrum (here, denoted as A–G; for details, see Fig. S1 or Table 2), resulting in 14 different LR models. These spectral ranges were recognized in our previous article [13] as highly informative (from the dating perspective) thanks to regularised MANOVA (rMANOVA). They consisted of time-dependent features characterized by b^2/w^2 above one, meaning that the variance between differently-aged bloodstains (b^2) was higher than within groups of bloodstains characterized by the same age (w^2).

Having established LR models, their validation was undertaken. For research purposes, such a procedure in which each G bloodstain samples from the first database (set No. 1) is compared with a single recovered sample from the second database was insufficient. Therefore, to validate the LR model, the procedure was repeated for G recovered samples stored in the second database (set No. 2). In this way, there were G^2 comparisons, of which G concerned comparing samples of the same age (i.e., $t_R = t_E$). Since the H_1 hypothesis should be supported in this case, and the expected LR value is above one,

false negatives appear when LR < 1. To estimate the false positive answers, where the H_2 should be supported (LR < 1), $\boldsymbol{G^2 - G}$ concerned comparing samples of different ages (i.e., $t_R \neq t_E$); thus, each LR > 1 is classified as the answer providing incorrect support for the H_1 (false positive). Both databases (sets 1 and 2) were used to test the performance of the LR models. Moreover, the false positives and negatives experiments were carried out ten times, each time generating a set to train and test the LR model. Therefore, the final error rates will be presented as an average of these partial results.

Additionally, the quantitative aspect of LR models' functioning, namely their ability to point out the strength of support for each hypothesis, was assessed using empirical cross entropy (ECE) [17, 45, 46]. The ECE is a performance metric that relies on assigning an appropriate penalty for each yielded LR value. The penalty magnifies with the increasing support for the incorrect hypothesis according to the logarithmic strictly proper scoring rules (a description of the ECE procedure for the performance assessment is presented in the Supplementary Material, SM). Consequently, the higher the support for the wrong hypothesis, the more severe the given error is because the greater penalty is assigned to the evaluated LR model.

All calculations were performed in R software [47] using scripts home-written by A. Martyna and available R packages.

3. Results and discussion

The following example may be considered to understand the comparison approach proposed in this study. Among the traces secured at the crime scene, there were bloodstains identified as originating from the suspect. The defense questioned the relevance of these findings by casting doubt on the timing of the bloodstains' deposition. Thus, the critical information for establishing the suspect's connection to the considered event was the time of the trace formation. Contrary to the conventional dating approach, obtaining this information does not have to come down to determining the absolute age of the sample by applying a previously established calibration model. It suffices to compare the degradation state of the evidential material with some reference bloodstains, created in such a way to match hypotheses, which represent the position of the prosecutor and the defense.

To obtain a well-balanced evaluation of the evidence, it would be desirable to create two different sets of reference materials, producing – at best – results of two different comparisons. These reference materials are obtained through supervised aging, simulating – as closely as possible – the actual settings of evidence decomposition at the crime scene. In other words, a similar volume of blood, freshly drawn from the donor (the same individual who was the donor of the questioned trace), should be deposited on an identical substrate and stored under similar conditions as those observed at the crime scene. If possible, one of these reference sets is then subjected to the degradation process for the time corresponding to the prosecution's version of events. The other should be aged according to the defense's scenario.

Having both evidential and reference materials, the proposed dating procedure enters the next phase, namely the characterization of the state of blood traces degradation. Finally, the resulting physicochemical data (e.g., spectral signatures) should be interpreted in the context of the prosecutor's and the defense's perspectives and communicated understandably to the justice system representatives. An approach founded on the likelihood ratio framework [16–18] that forms the proposed comparative procedure's backbone is recommended to achieve this goal.

3.1. Preprocessing of Raman signatures using genetic algorithm (GA)

As well-documented, the correct choice of preprocessing usually improves the performance of statistical models, including LR models, which – at the same time – can be significantly impaired in case of improperly handled signals. Altogether, adequate preprocessing is a crucial step in data treatment. In the case of Raman spectroscopy-derived signals, usually, two basic preprocessing steps are vitally crucial for feasible analysis. The first is baseline correction methods to remove the baseline effects arising, among others, from fluorescence and other additive features in the spectra. The second group consists of normalization procedures whose task is to remove multiplicative effects related, for instance, to laser intensity fluctuations or out-of-focus contributions. Thus, normalization usually boils down to multiplying the signal by a scaling value to make the corresponding intensities, which, in theory, should not pose any differences as comparable across spectra as possible. Unfortunately, regardless of the final goal of data modeling, the difficulty here is that there is no universal combination of preprocessing methods suitable for each specific analysis's objective and the type of spectral distortions.

In order to address this problem, before developing LR models, the strategy based on applying the genetic algorithm (GA) to indicate the combination of preprocessing tools – optimal given the study purpose – was implemented. Table 2 collects all the preprocessing solutions suggested by the GA as the most effective combinations in yielding the highest between-source and within-source variance (b^2/w^2) ratio for the first latent variable computed from rMANOVA as a quality parameter. In such a way, as broadly discussed in the study of Martyna [28], the discrimination of signals corresponding to differently-aged bloodstains was significantly enhanced, as the preprocessing procedure ensured the best exposition of differences between sources while minimizing the casual variations within sources. The suitability of the proposed methodology was ultimately verified by the performance of developed LR models built for the first rMANOVA's eigenvector, which were reflected in – among others – the levels of false positive and false negative rates (shown as boxplots in Fig. 4) and also numerically summarized in Table S2.

Table 2 Details of preprocessing strategies applied when sets No. 1 and No. 2 were training sets found using a genetic algorithm (GA). All abbreviated methods are explained in Table S1 (SM) and partially in section *2.2.1*.

Spectral range [cm^{-1}]	Reference: set No. 1	Reference: set No. 2
A: 1200-1410	**Denoising**: SG, polynomial degree p = 3 **Baseline correction**: SNIP, clipping window w = 26 **Normalization**: PQN	**Denoising**: DWT, Daubechies Least Asymmetric 4, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: RBE, h = 0.4, b = 2 **Normalizatio**n: PQN
B: 1200-1300; 1500-1700	**Denoising**: DWT, Daubechies Least Asymmetric 8, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: CWTAsWPLS, m = 2, $\lambda = 1 \cdot 10^8$ **Normalization**: SNV	**Denoising**: DWT, Coiflets 1, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: polyQR, polynomial degree p = 5, q = 0.05 **Normalization**: SNV
C: 1200-1300; 950-1020	**Denoising**: DWT, Coiflets 5, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: RBE, h = 0.4, b = 2.2 **Normalization**: PQN	**Denoising**: DWT, Daubechies Least Asymmetric 4, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: RBE, h = 0.3, b = 2.5 **Normalization**: PQN
D: 1200-1410; 950-1020	**Denoising**: DWT, Coiflets 5, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: multiWAsPLS, m = 2, λ = 10, $\mu = 1 \cdot 10^9$ **Normalization**: SNV	**Denoising**: DWT, Coiflets 1, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: polyQR, polynomial degree p = 5, q = 0.05 **Normalization**: PQN
E: 1200-1410; 1500-1700	**Denoising**: SG, polynomial degree p = 3 **Baseline correction**: polyQR, polynomial degree p = 6, q = 0.01 **Normalization**: SNV	**Denoising**: SG, polynomial degree p = 5 **Baseline correction** pWAsPLS, m = 2, $\lambda = 1 \cdot 10^6$, w = 0.001 **Normalization**: SNV
F: 1200-1300; 1500-1700; 950-1020;	**Denoising**: DWT, Daubechies Least Asymmetric 4, d = 10, t = universal, c = soft, sd = mad **Baseline correction**: CWTAsWPLS, m = 2, $\lambda = 2 \cdot 10^8$ **Normalization**: SNV	**Denoising**: DWT, Daubechies Least Asymmetric 8, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: SNIP, clipping window w = 30 **Normalization**: SNV
G: 1200-1410; 1500-1700; 950-1020;	**Denoising**: SG, polynomial degree p = 4 **Baseline correction**: SNIP, clipping window w = 26 **Normalization**: SNV	**Denoising**: DWT, Coiflets 1, d = 10, t = SURE, c = soft, sd = mad **Baseline correction**: multiWAsPLS, m = 2, λ = 10, $\mu = 1 \cdot 10^8$ **Normalization**: SNV

d – decomposition level for denoising, t – threshold estimation, c – thresholding policy, sd – dispersion estimate, m – order of differences, λ – penalty, μ – penalty term, h – the proportion of signal points for the local regression, b – robustness parameter, q – quantile, w – weights

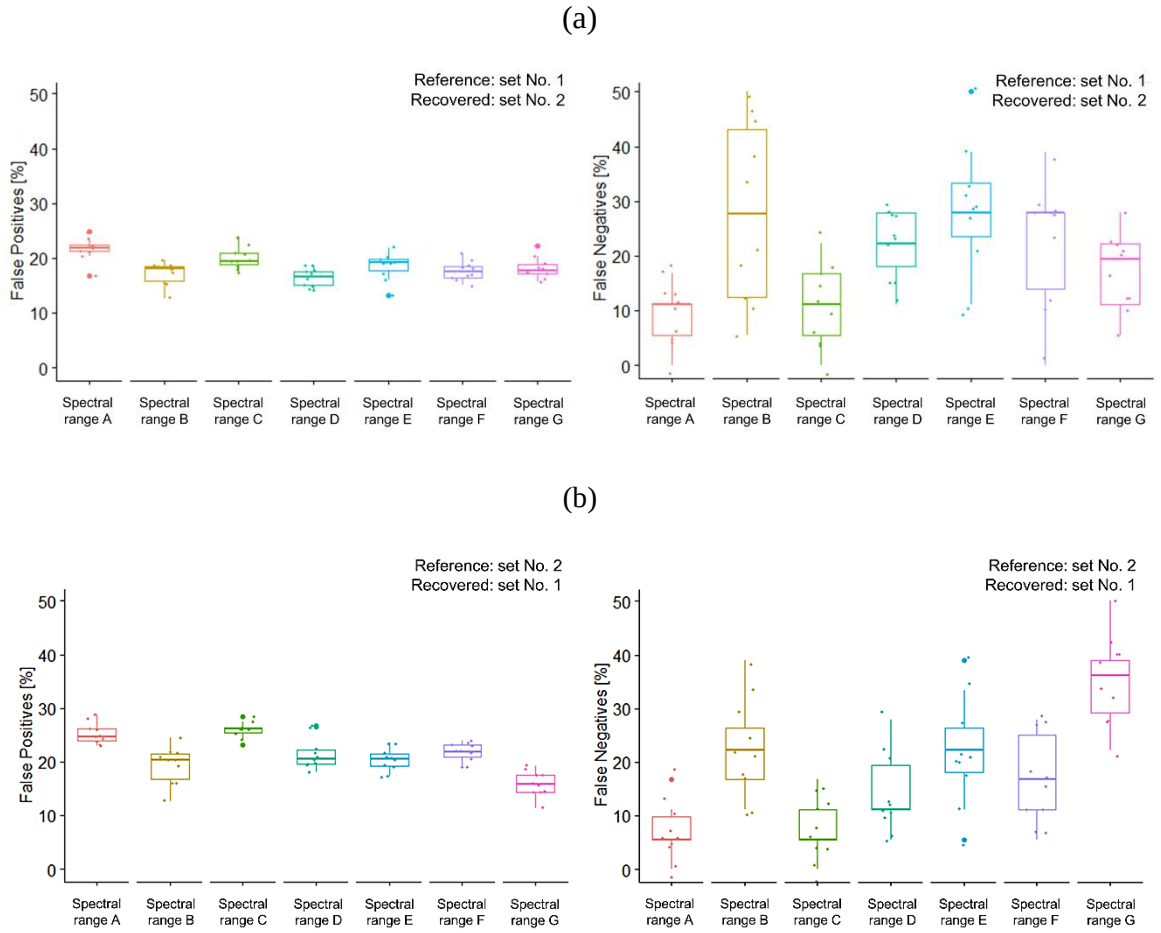


Fig. 4. The levels of false positive and false negative responses of LR models constructed for Raman spectra of bloodstains prepared using set No. 1 (a) and set No. 2 (b), respectively, for training the LR models.

The selection of preprocessing approach was individually determined by the type of distortion present in the recorded signals. As shown in Table 2, GA selected a different combination of preprocessing methods for almost each developed LR model, ensuring their robustness. One can also notice that denoising and baseline correction methods practically did not indicate the superiority of any specific methods with respect to others. Just a slight trend is observed among the normalization methods. The false negative responses (Fig. 4) observed for individual LR models revealed that the lowest error rates were associated with models developed on the following spectral ranges: 1200–1410 cm^{-1} and 1200–1300 cm^{-1} combined with 950–1020 cm^{-1}. This observation is also supported by the empirical cross entropy (ECE) plots in Fig. 5 (ECE plots corresponding to remaining LR models are depicted in Fig. S2 and S3). LR models based on these two spectral ranges yield incomparably more satisfying ECE outcomes. In both cases (denoted as spectral ranges A and C in Table 2), PQN – not SNV – was used as the normalization method. The most logical conclusion would attribute this model performance difference to the varying spectral ranges used to construct the LR models. In other words, variants A and C, covering the ranges 1200–1410 cm^{-1} and 1200–1300 cm^{-1} combined with 950–1020 cm^{-1}, respectively, are characterized by bands that exhibit the most pronounced time-related variations (high

values of b^2/w^2 ratios). Changes in spectral characteristics, reflected mainly in such spectral features as hemoglobin aggregation (1255 cm^{-1} and 976 cm^{-1}) and oxidation markers (between 1300–1400 cm^{-1}), are well-recognized effects of the formation of Hb degradation products [8–13]. This also confirms the results from our previous study [13], where it was found that the aging of blood traces boils down primarily to the conversion of oxyHb to metHb and HC, followed by the aggregation of heme species.

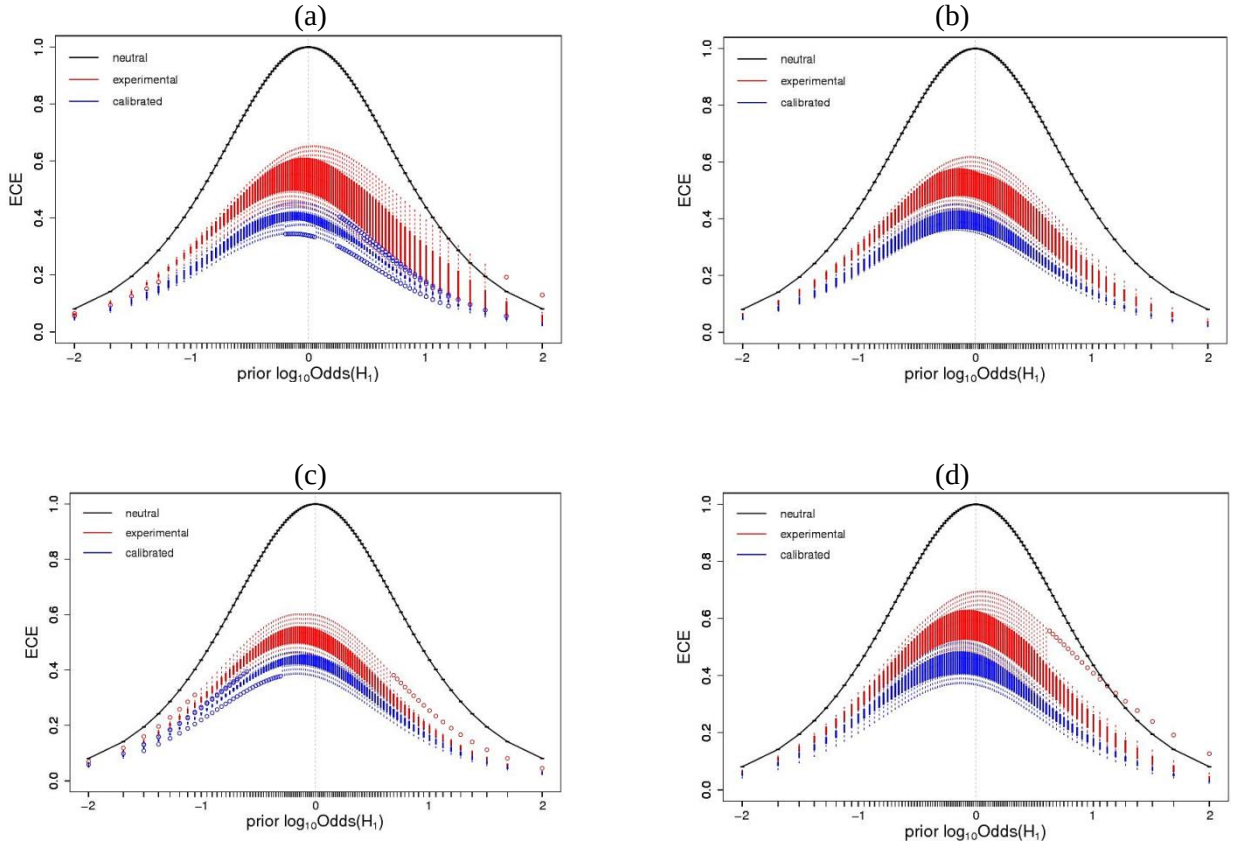


Fig. 5. Box-empirical cross entropy plots for the LR models constructed for Raman spectra of bloodstains prepared using the preprocessing strategies indicated by the genetic algorithm (GA). Two depicted LR models were trained on set No. 1 and tested using set No. 2 for the following spectral ranges: 1200–1410 cm^{-1} (a); 1200–1300 cm^{-1}, and 950–1020 cm^{-1} (b). The two remaining LR models were trained on set No. 2 and tested using set No. 1 for the following spectral ranges: 1200–1410 cm^{-1} (c); 1200–1300 cm^{-1}, and 950–1020 cm^{-1} (d).

3.1.2 Performance of the likelihood ratio models

Figure 4 and Table S2 portray the overall performance of the LR models under investigation, computed using the first (set No. 1) and the second (set No. 2) databases. The general findings expose that false positive (F_P) rates obtained for each of the seven model variants are somewhat more comparable to each other than the levels of false negative (F_N) answers. F_P oscillated around 20%, and the dispersion of results within each model is not as evident as in F_N.

This observation can be explained by the higher similarity of compared bloodstain samples at more advanced stages of degradation. Inconsiderable differences between the spectral characteristics of blood traces, a consequence of a slow-down in the aging process, appear to equally affect the LR models'

capability to distinguish between differently-aged samples, leading to the increase in the levels of F_P answers. The observed misleading support for the hypothesis about the common source of bloodstains (H_1: the recovered and reference bloodstains are characterized by the same TSD) obviously cannot be reduced by implementing other preprocessing techniques or signal manipulation in general. This is because the factor determining the F_P answers here is the rate of blood degradation itself and the ability to monitor it with a given analytical technique. Therefore, it needs to be clearly stated that the purpose of the present research was not to demonstrate the superiority of Raman spectroscopy over any other analytical method capable of characterizing the degradation state of bloodstains. Its objective was solely to evaluate the performance of the novel dating framework, which – eventually – proved effective when combined with Raman spectroscopy, especially at the initial stage of aging (as depicted in Fig. 6).

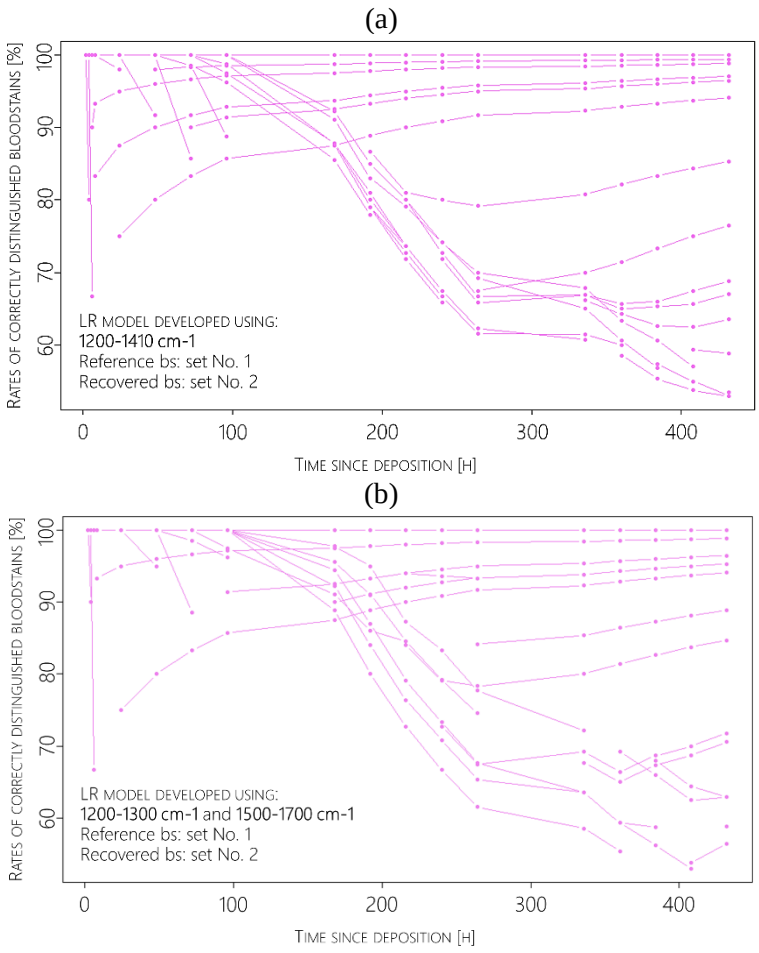


Fig. 6. The percentage of correctly distinguished differently-aged blood traces in the function of aging time [h] obtained when using LR models developed for spectral ranges A (a) and B (b). In both cases, the reference and recovered bloodstains originated from sets No. 1 and No. 2, respectively.

Reporting F_P responses over the entire analyzed aging period may be misleading due to its low information content. It should be clear that the longer the degradation time, the higher similarity of spectroscopic signals; hence the higher percentage of incorrect answers will be obtained. Consequently, it is much more reasonable to present the F_P rates concerning the time elapsed since bloodstains

deposition, as it would better report on the effectiveness of the dating approach. An example of such dependency for LR models developed for spectral ranges A and B is shown in Figures 6a and 6b, respectively. In both presented cases, LR models were trained using set No. 1. Instead of F_P responses, the plots depict the percentage of correctly distinguished differently-aged blood traces plotted against the aging time. Thus, one needs to subtract the value shown in the plot for a given time point from 100% to obtain the percentage of false positives. For example, for the first time point considered (two hours elapsed since bloodstain formation), the average percentage of correct discrimination is 98.82%. Thus, the corresponding F_P value corresponds to 1.18%. What is more, the primary purpose of Figure 6 is to demonstrate some time trends characterizing the performance of developed LR models. This time trend indicates nearly 100% discrimination effectiveness of the LR models within the first seven days of degradation, decreasing after about 170 hours, which is a benchmark performance. As can be seen, even though the overall performance of these models is entirely different (due to F_N rates, 9.45% and 27.77% for models A and B, respectively), their ability to discriminate between differently-aged samples over time correctly – reflected in the rates of F_P responses – is exceedingly high and relatively similar.

Consequently, the factor that actually differentiates the developed LR models concerning their effectiveness in solving the considered comparison problem is the already mentioned reduction of within-source variations, reflected in the percentage of false negative answers. After inspection of F_N rates, it becomes clear that LR models based on PQN-normalized spectral ranges A (1200–1410 cm^{-1}) and C (1200–1300 cm^{-1} combined with 950–1020 cm^{-1}) outperform the remaining LR models. F_N rates for these two models' variants (bolded in Table S2) are less dispersed and oscillate around 10% while they reach ca. 30% for the remaining spectral ranges. Thus, once again, it turns out that from the dating perspective, the most informative Raman features are the Hb aggregation markers (e.g., 1255 cm^{-1} or 976 cm^{-1}) and so-called oxidation markers located between 1300–1400 cm^{-1} [48–50]. Interestingly, the greater the influence of the 1500–1700 cm^{-1} range in the developed model, the worse its performance. An example of this detrimental effect of including the 1500–1700 cm^{-1} range in the model is the one based on spectral range B, namely 1200–1300 cm^{-1} combined with 1500–1700 cm^{-1}. This may suggest that bands appearing in this range, which contains the core-size and spin state markers [13], are more susceptible to random variation (e.g., resulting from dietary factors) than changes due to bloodstains aging. Hence, future studies should primarily focus on identifying factors influencing the degradation kinetics to simplify the process of supervised aging.

Finally, the ECE plots confirmed the conclusions reached upon examination of the levels of F_N answers, proving that LR models based on spectral ranges A (1200–1410 cm^{-1}) and C (1200–1300 cm^{-1} combined with 950–1020 cm^{-1}) produce the most desirable performances. The ECE plots corresponding to those two LR models are presented in Fig. 5. The plots for the five remaining spectral ranges (B and D–G) are provided in SM (Fig. S2 and S3). The experimental and calibrated curves (indicated in red and blue, respectively) are represented by boxplots positioned at the considered prior odds (prior probability quotients $Pr(H_1)/Pr(H_2)$). Each boxplot accounts for the ECE values calculated for a given

prior odds using the available likelihood ratio values. For the best-performing models, the reduction of information loss (or gain of information) reaches even ca. 50% (e.g., Fig. 5c), which is more than an acceptable outcome given the small databases. In other words, it means that after the bloodstain evidence analysis using developed LR models, a relatively large amount of information concerning the uncertainty about the correct hypotheses was explained. The ECE values also evidenced that the models based on 1500–1700 cm^{-1}-containing regions (Fig. S2 and S3) represent the worst solutions indisputably. Not only did they deliver the highest rates of false negative answers, but they also yielded incomparably inferior ECE outcomes.

It should also be noted that in the case of poorly performing models, the ECE plots deviated from the desired bell-shaped curves. The ECE values went beyond the neutral curve, especially for the positive logarithm of the prior odds, i.e., $\log_{10}\text{Odds}(H_1) > 0$. This deterioration may be caused by only a single sample delivering strong misleading support for the incorrect hypothesis (in this case – H_1). In particular, when small databases are considered, the penalty assigned to a misleading LR value might outweigh the importance given to the LRs supporting the correct hypothesis. For this reason, these inferior ECE results are not necessarily evidence of poorly performing LR models. The ECE approach may simply not be the most robust performance measure. Thus, it should also be interesting to explore that issue further in the future, trying to verify whether the obtained poor LR model performance, assessed using the ECE approach, may be caused by just a single misleading value. Moreover, some differences between the experimental and calibrated ECE plots obtained even for well-performing models (Fig. 5) hold promise for improving the models' effectiveness. The simplest way to refine the model may be to expand its databases to capture better all the relevant features characteristic of the whole population of the analyzed bloodstains.

4. Conclusions and future perspectives

The objective of the presented research was to verify the possibility of estimating the age of blood deposits – characterized by Raman signatures – through their comparison with "artificially" created reference materials. Thus, the conventional quest for the bloodstains dating method, based on regression analysis, was replaced with a discrimination problem, solvable through the likelihood ratio approach.

Given the inherent variability of analyzed materials and the multidimensionality of the considered issue, the capability of established models to discriminate between differently-aged samples was the benchmark performance, providing preliminary evidence of the effectiveness of the proposed dating framework. The best models, founded on 1200–1410 cm^{-1} and 1200–1300 cm^{-1} combined with 950–1020 cm^{-1} ranges, correctly distinguished nearly 100% of differently-aged blood traces within the first week of degradation. Additionally, across the monitored time range, these models delivered approximately 10% of false negatives and 20% false positive answers, while the ECE plots evidenced ca. 50% reduction in information loss. The false negative rates were gratifying, while the decent ECE

results, at this point, should not be too much of a concern. As already mentioned, when combined with scarce databases, ECE often leads to misleading conclusions due to its sensitivity to individual LR values supporting the incorrect hypothesis. In contrast, the 20% false positives may not be entirely satisfactory. Unfortunately, the level of false positive answers is not the product of improper preprocessing or data modeling but rather an effect of a significant slowdown in bloodstain aging processes that could no longer be effectively monitored by Raman spectroscopy. Hence, it should be concluded that Raman spectroscopy cannot be considered a panacea for tracing blood degradation processes and, in fact, was never considered one. It might be unrealistic to expect a single method to meet analytical requirements posed by any possible evidential bloodstain. In this study, Raman spectroscopy was only used to verify the effectiveness of LR models designed to discriminate differently-aged blood traces. This effectiveness – still very high during the first week of degradation (nearly 100% correct discriminations) – will increase once other analytical methods are applied, better suited to tracking aging changes in bloodstains. Thus, an interesting way forward may be inviting other researchers for the joint development of this approach by implementing other analytical techniques capable of probing the state of bloodstains degradation over extended periods (e.g., FT-IR, UV-Vis spectroscopies, or even RNA analysis), which is now the subject of our further research.

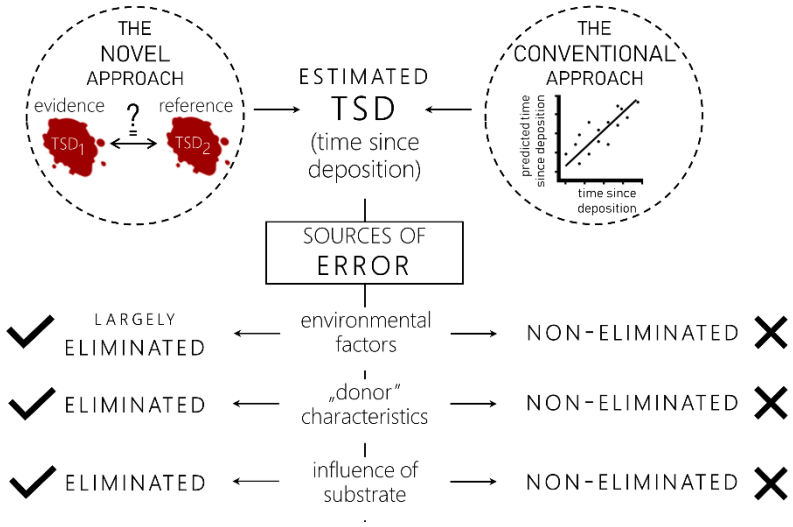


Fig. 7. The novel approach *versus* the conventional dating approach – sources of error.

Another important note for future studies is that the high reproducibility of the supervised aging procedure is the determining factor for reliable and valid dating. In other words, a similar volume of blood, freshly drawn from the donor (if possible, the same individual who was the donor of the questioned trace), should be deposited on an identical substrate and stored under similar conditions as those observed at the crime scene. To reach valid TSD estimates, bloodstains serving as reference samples for building models must replicate the aging process of the evidence as accurately as possible. Irrespective of the monitoring method applied, the bloodstains' resultant aging profile depends on the aging kinetics, which is affected by such influential factors as the substrate, environmental conditions

persisting during aging (controlled by climatic chambers), and the initial composition of the bloodstain. Therefore, the success of the proposed approach will depend on how much the above factors influence degradation and, above all, on how much it is possible to control them during supervised aging. This way, sources of error in TSD estimates that occur during the conventional dating approach would be largely eliminated (Fig. 7).

Indeed, the mere statement that errors arising from external factors will be largely eliminated is not a sufficient guarantee of the reliability of the proposed methodology. Hence, once again, it should be reminded that to meet the ISO/IEC 17025 international quality standards [16] or – simply – to ensure the validity of any novel methodology [51–53], a forensic laboratory must undertake a comprehensive evaluation of all factors that could impact the examination result during the validation process of a method intended for use in casework. In this specific case, after identifying the factors (and their combinations) that influence the degradation of bloodstains, the accuracy of TSD estimates using the new methodology can be fully assessed. It will be achieved by validating likelihood ratio models, including determining the rates of false positive and false negative responses, for "evidential" bloodstains degrading under different conditions, not only those presented in the present study. Only then will it be possible to consider advancing this methodology to practical application.

Obviously, it would be somewhat naïve to assume the possibility of recreating precisely the same conditions as evidence degradation. Nonetheless, recreating them as closely as possible to those prevailing during the degradation of the evidential bloodstain may be sufficient to answer the prosecutor's questions regarding the time of traces' formation. Thus, in the first place, future research should focus on identifying factors that may affect the bloodstains' degradation – using experimental design methods [54] – so that the supervised aging procedures focus on controlling only those specific factors. One of the most significant and likely most challenging tasks will be understanding the relationship between the initial blood composition (the "donor" characteristics) and the kinetics of its degradation. Determining this relationship will allow answering whether the supervised aging procedure will need to be carried out using blood from the person of interest/suspect or whether a comparison sample could be constructed from any blood. In the latter case, the proposed dating procedure would still be feasible, especially if the suspect was actually guilty of the alleged crime and refused to provide a blood sample for testing. In a much more likely scenario, when the source of blood will determine the rate of aging, it is worth remembering that forensic science can be used not only to prove the guilt but also to exonerate innocent people from the charges brought against them. In the era of social doubts about the reliability of forensic sciences, it is important to avoid jailing people because of flawed or insufficient forensic techniques. The inability to estimate the time of bloodstain formation creates an analytical gap and carries the risk of convicting an innocent person. In other words, in cases where a blood sample is to be obtained from the wrongly accused person and could ultimately lead to one's acquittal, the problem of obtaining blood for comparative analysis should disappear.

Finally, it is important expressly to highlight that the case-by-case approach founded on the idea of comparing the evidence with some reference materials should be of a wider forensic application. The proposed framework is expected to enrich the forensic arsenal of analytical tools when answering the age of other evidence, such as different types of body fluids, organic gunshot residues [55], or fingerprints [56]. In this regard, this novel dating approach is quite universal. It just requires to be complemented with an analytical technique adequate for characterizing aging profiles of the forensic trace of interest, while the fundamental idea of dating – founded on the LR-based comparison – remains the same.

Funding

The research did not recive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest

There are no conflicts to declare.

References

1 C. Weyermann, O. Ribaux, *Sci. Justice*, 2012, **52**, 68–75.
2 O. Ribaux, S. Walsh, P. Margot, *Forensic Sci. Int.*, 2006, **156**, 171–181.
3 O. Ribaux, A. Baylon, C. Roux, O. Delémont, E. Lock, C. Zingg, P. Margot *Forensic Sci. Int.*, 2010, **195**, 10–16.
4 L. Tomellini, *Arch. d'Antropologie criminelle de Criminol.*, 1907, **14**, 2.
5 R.H. Bremmer, K.G. de Bruin, M.J.C. van Gemert, T.G. van Leeuwen, M.C.G. Aalders, *Forensic Sci. Int.*, 2012, **216**, 1–11.
6 G. Zadora, A. Menżyk, *TrAC*, 2018, **105**, 137–165.
7 V. Sharma, R. Kumar, *TrAC*, 2018, **107**, 181–195.
8 K. M. Marzec, A. Rygula, B. R. Wood, S. Chlopicki, M. Baranska, *J. Raman Spectrosc.*, 2014, **46**, 76–83.
9 P. Lemler, W.R. Premasiri, A. DelMonaco, L.D. Ziegler, *Anal. Bioanal. Chem.*, 2014, **406**, 193–200.
10 K.C. Doty, G. McLaughlin, I.K. Lednev, *Anal. Bioanal. Chem.*, 2016, **408**, 3993–4001.
11 K. C. Doty, C. K. Muro, I. K. Lednev, *Forensic Chem.*, 2017, **5**, 1–7.
12 H.J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N.J. Fullwood, B. Gardner, P.L. Martin-Hirsch, M.J. Walsh, M.R. McAinsh, N. Stone, F.L. Martin, *Nat. Protoc.*, 2016, **11**, 664–687.
13 A. Menżyk, A. Damin, A. Martyna, E. Alladio, M. Vincenti, G. Martra, G. Zadora, *Talanta*, 2020, **209**, 120565.
14 H. Lin, Y. Zhang, Q. Wang, B. Li, P. Huang, Z. Wang, *Sci Rep.*, 2017, **7**, 13254.
15 R. Kumar, K. Sharma, V. Sharma, *Sci. Justice*, 2020, **60**, 538–546.
16 C. Gannicliffe, End User Commentary on Investigating the Age of Blood Traces: How Close Are We to Finding the Holy Grail of Forensic Science?, in: S. Francese (Ed.), Emerging Technologies for the Analysis of Forensic Traces, *Springer*, 2019, pp. 129–132.

17 G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, *Wiley*, Chichester, 2014.
18 ENFSI Guideline for Evaluative Reporting in Forensic Science. Strengthening the Evaluation of Forensic Results across Europe. European Network of Forensic Science Institutes, 2015, https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf, (accessed 31st August 2021).
19 A. Martyna, G. Zadora, Hybrid Likelihood Ratio Models for Forensic Applications: a Novel Solution to Determine the Evidential Value of Physicochemical Data, in: L. Dennany (Ed.), Challenges in Detection Approaches for Forensic Science, *RSC*, 2021, pp. 198–231.
20 A. Damin, et al., European patent No. WO2017077513 (A1), (2017).
21 M. Signorile, F. Bonino, A. Damin, S. Bordiga, *Top. Catal.*, 2018, **61**, 1491–1498.
22 A. Martyna, G. Zadora, T. Neocleous, A. Michalska, N. Dean, *Anal.Chim. Acta*, 2016, **931**, 34–46.
23 A. Martyna, K.-E. Sjastad, G. Zadora, D. Ramos, *Talanta*, 2013, **105**, 158–166.
24 A. Martyna, A. Michalska, G. Zadora, *Anal. Bioanal. Chem.*, 2015, **407**, 3357–3376.
25 A. Michalska, A. Martyna, J. Zieba-Palus, G. Zadora, *J. Raman Spectrosc.*, 2015, **46**, 772–783.
26 A. Martyna, G. Zadora, D. Ramos, *J. Anal. Appl. Pyr.*, 2018, **133**, 198–215.
27 J. Engel, L. Blanchet, B. Bloemen, L. Heuvel, U. Engelke, R. Wevers, L. Buydens, *Anal. Chim. Acta*, 2015, **89**, 1–12.
28 A. Martyna, A. Menżyk, A. Damin, A. Michalska, G. Martra, E. Alladio, G. Zadora, *Chemom. Intell. Lab. Syst.* 2020, **202**, 104029.
29 O. Kvalheim, F. Brakstad, Y.-Z. Liang, *Anal. Chem.*, 1994, **66**, 43–51.
30 S. Mallat, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1989, **7**, 674–693.
31 A. Savitzky, M. Golay, *Anal. Chem.*, 1964, **36**, 1627–1639.
32 P. Eilers, *Anal. Chem.*, 2003, **75**, 3631–3636.
33 Z. Zhang, S. Chen, Y. Liang, *Analyst*, 2010, **135**, 1138–1146.
34 . Zhang, S. Chen, Y. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, H. Zhou, *J. Raman Spectrosc.*, 2010, **41**, 659–669.
35 P. Cadusch, M. Hlaing, S. Wade, S. McArthur, P. Stoddart, *J. Raman Spectrosc.*, 2013, **44**, 1587–1595.
36 J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan, *Anal. Chim. Acta*, 2010, **683**, 63–68.
37 A. Ruckstuhl, M. Jacobson, R. Field, J. Dodd, *J. Quant. Spectrosc. Radiat. Transf.*, 2001, **68**, 179–193.
38 C. Ryan, E. Clayton, W. Griffin, S. Sie, D. Cousens, *Nucl. Instrum. Methods Phys. Res. B*, 1988, **34**, 396–402.
39 C. Lieber, A. Mahadevan-Jansen, *Appl. Spectrosc.*, 2003, **57**, 1363–1367.
40 J. Zhao, H. Lui, D. McLean, H. Zeng, *Appl. Spectrosc.*, 2007, **61**, 1225–1232.
41 L. Komsta, *Anal. Bioanal. Chem.*, 2014, **406**, 1985–1998.
42 R. J. Barnes, M. S. Dhanoa, S. J. Lister, *Appl. Spectrosc.*, 1989, **43**, 772–777.
43 F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, *Anal. Chem.*, 2006, **78**, 281–4290.
44 B. W. Silverman. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, UK, 1986.
45 D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, *J. Forensic Sci.*, 2013, **58**, 1503–1518.
46 R. Haraksim, D. Ramos, D. Meuwly and C. E. H. Berger, *Forensic Sci. Int.*, 2015, **249**, 123–132.
47 R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018. URL, https://www.R-project.org/.
48 B. R. Wood, L. Hammer, D. McNaughton, *Vib. Spectrosc.*, 2005, **38**, 71–78.
49 B. R. Wood, L. Hammer, L. Davis, D. McNaughton, *J. Biomed. Opt.*, 2005, **10**, 14005–14013.

50 B. R. Wood, P. Caspers, G. J. Puppels, S. Pandiancherri, D. McNaughton, *Anal. Bioanal. Chem.*, 2007, **387**, 1691–1703.
51 A. Sikirzhytskaya, V. Sikirzhytski, L. Perez-Almodovar, I. K. Lednev, *Forensic Chem.*, 2023, **32**, 100468.
52 N. A. Nichols, I. K. Lednev, *Am. J. Anal. Chem.*, 2022, **13**, 1–8.
53 Y. V. Kistenev, A. V. Borisov, A. A. Samarinova, S. Colon-Rodriguez, I. K. Lednev, *Sci. Rep.*, 2023, **13**, 5384.
54 R. Leardi, *Anal. Chim. Acta*, 2009, **652**, 161–172.
55 S. Charles, N. Geusens, E. Vergalito, B. Nys, *Forensic Sci. Int.: Synergy*, 2020, **2**, 416–428.
56 K. Wertheim, *J. Forensic Identif.*, 2003, **53**, 42–49.

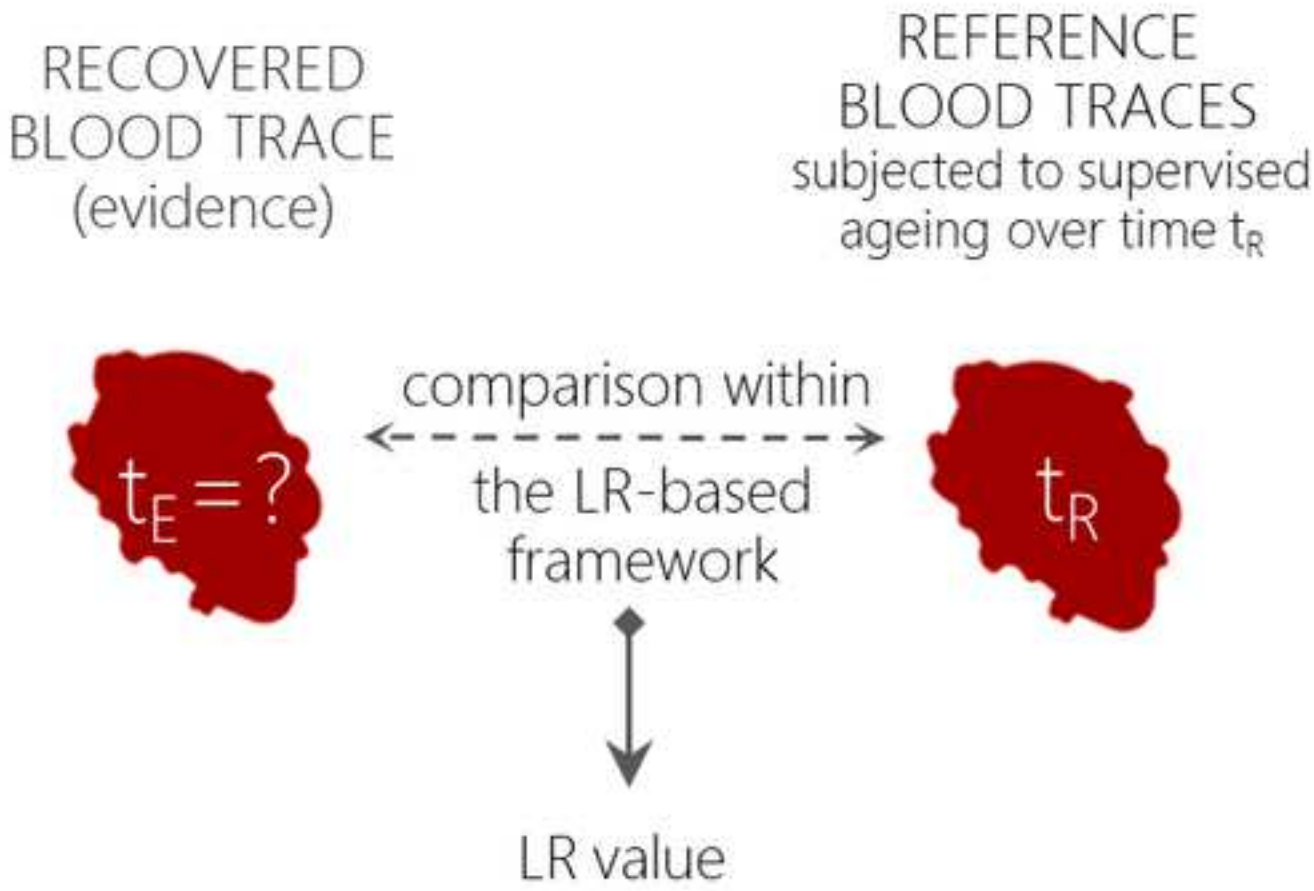


Figure 2

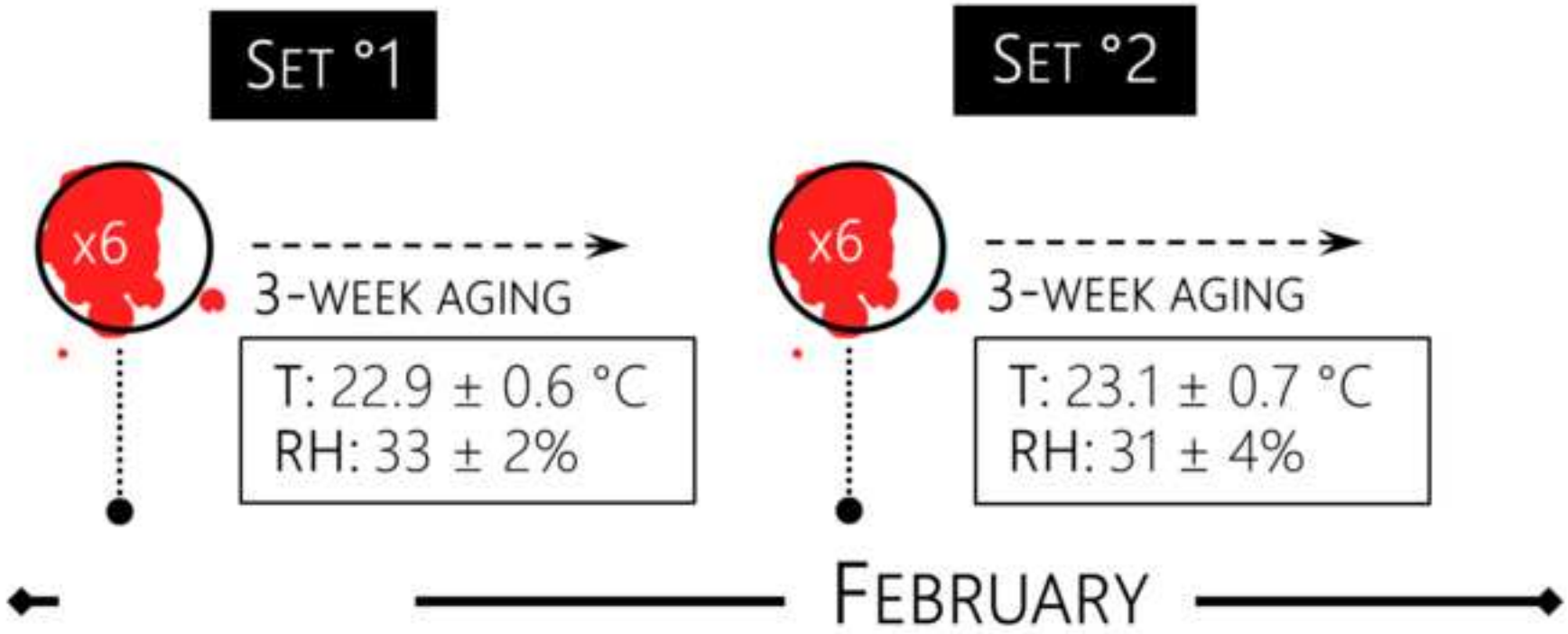


Figure 3

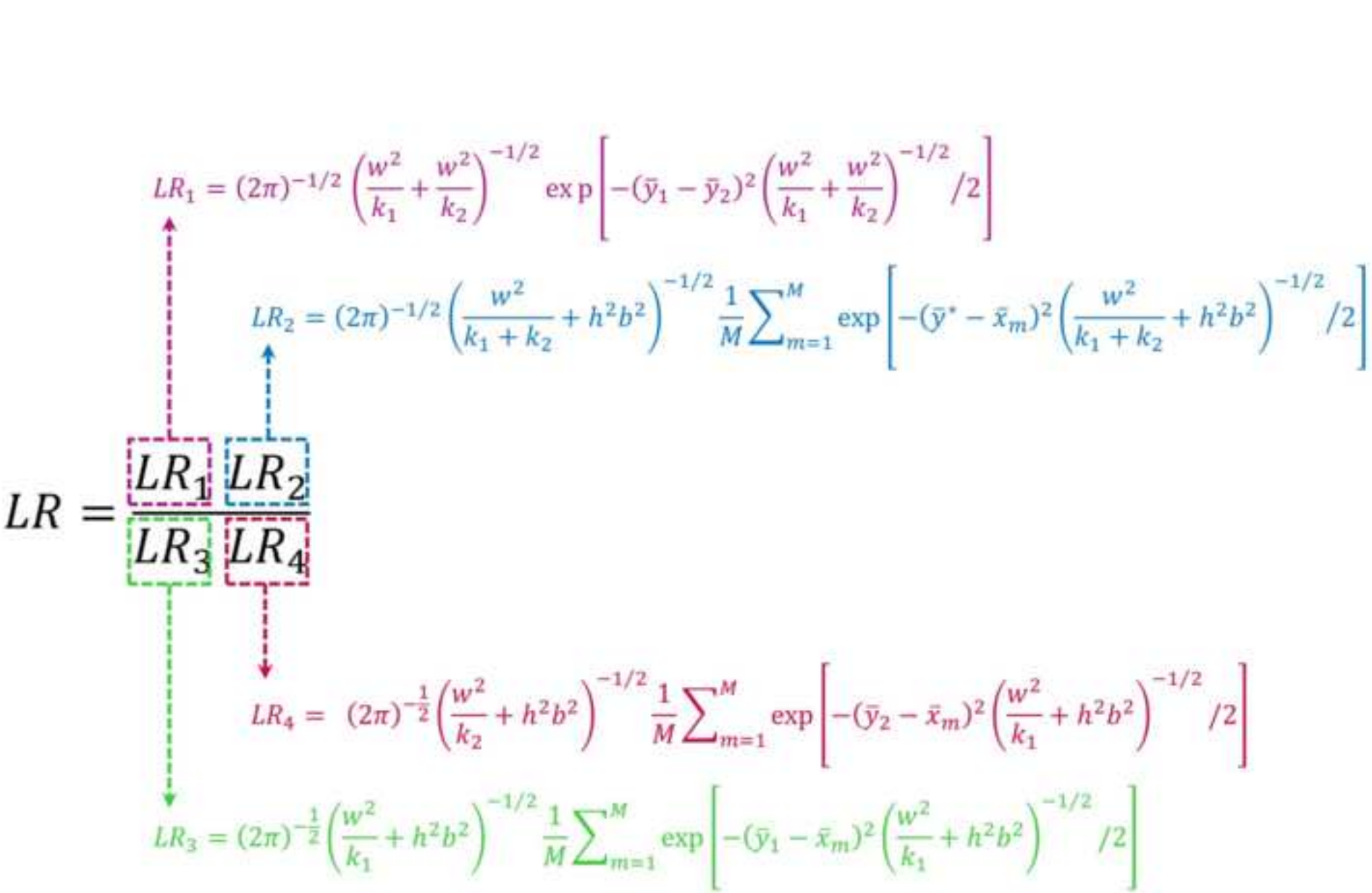


Figure 4a

Click here to access/download;Figure;Fig.4a.tif

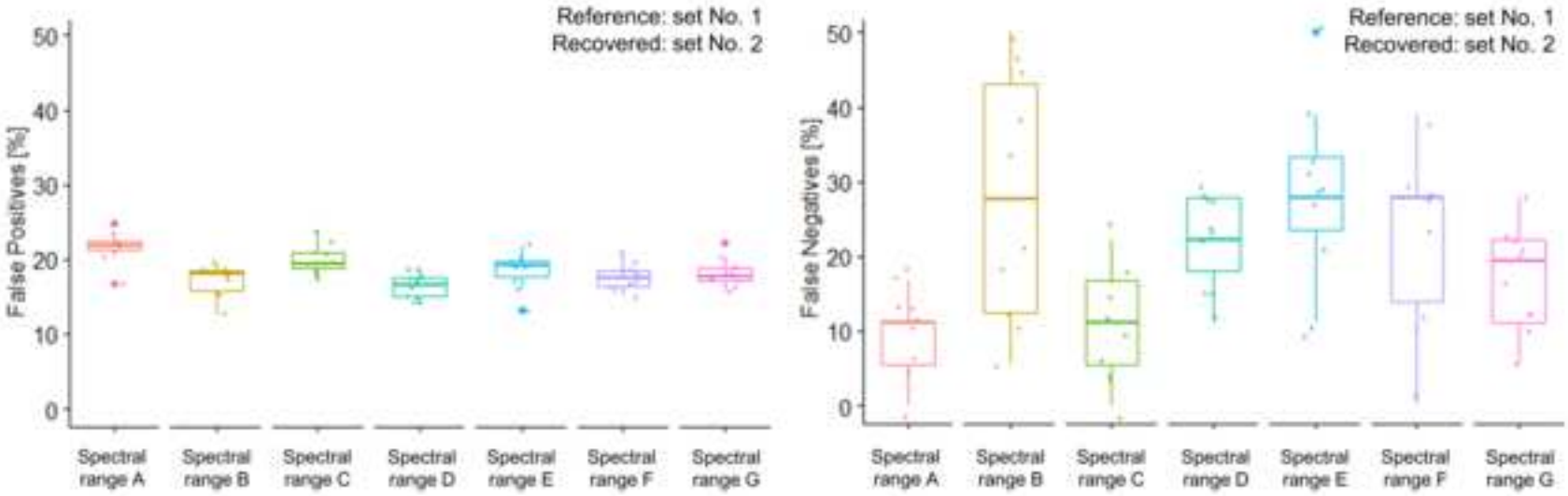


Figure 4b

Click here to access/download;Figure;Fig.4b.tif

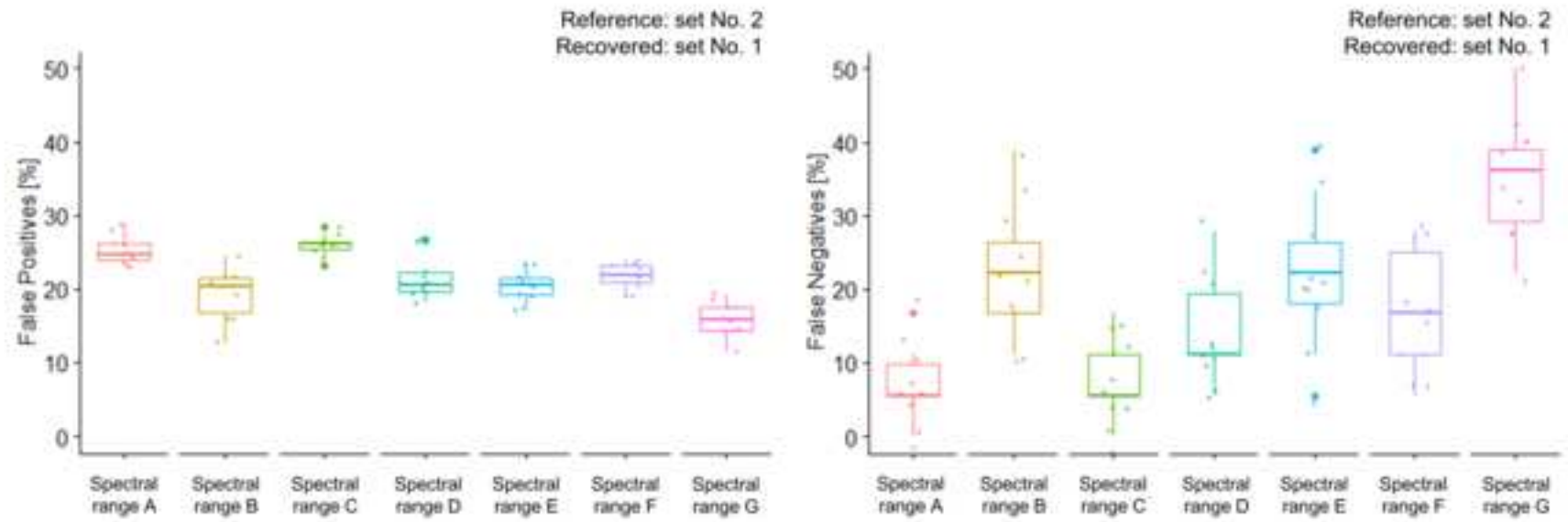


Figure 5a

Click here to access/download;Figure;Fig.5a.tif

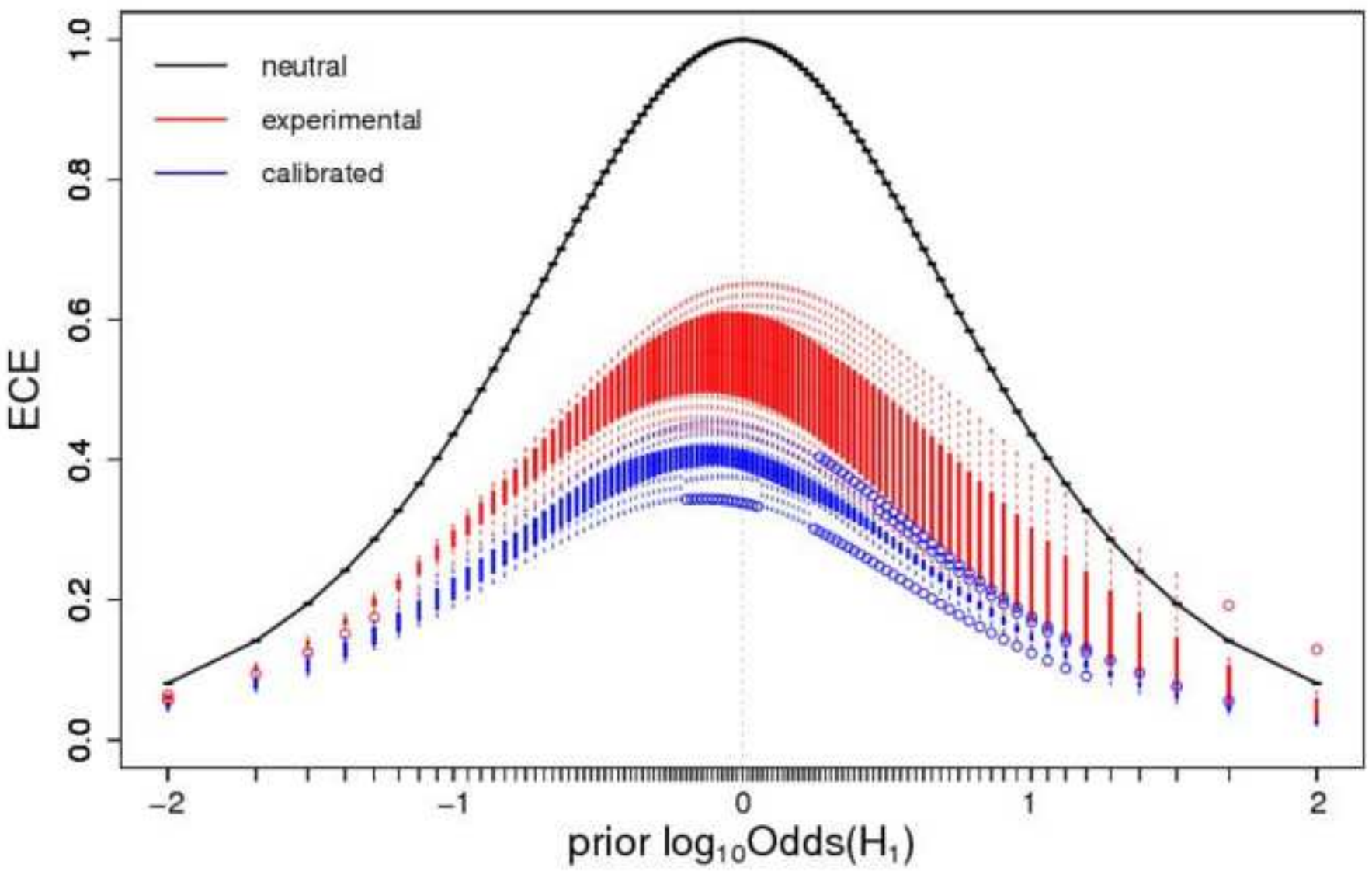


Figure 5b

Click here to access/download;Figure;Fig.5b.tif

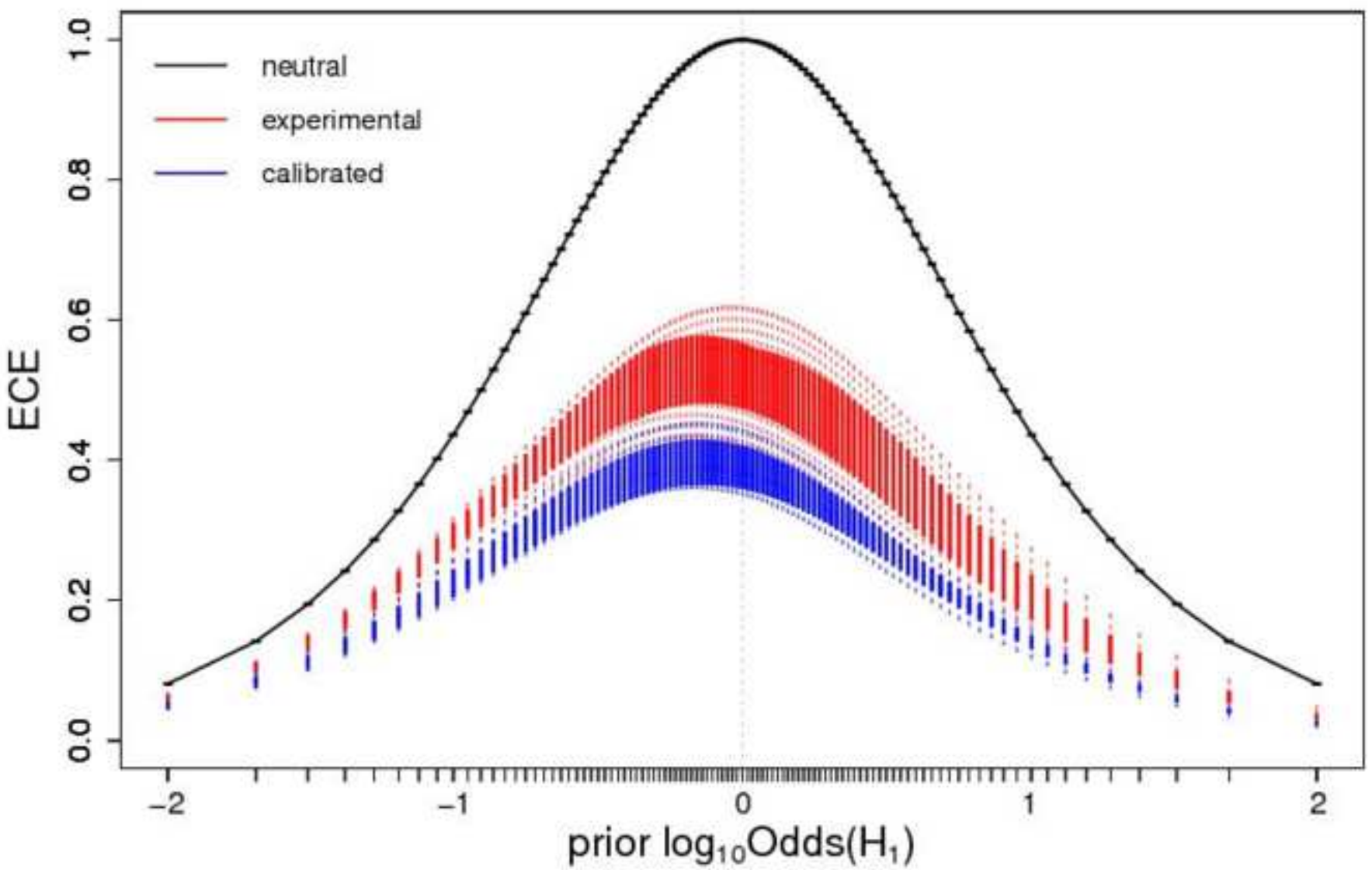


Figure 5c

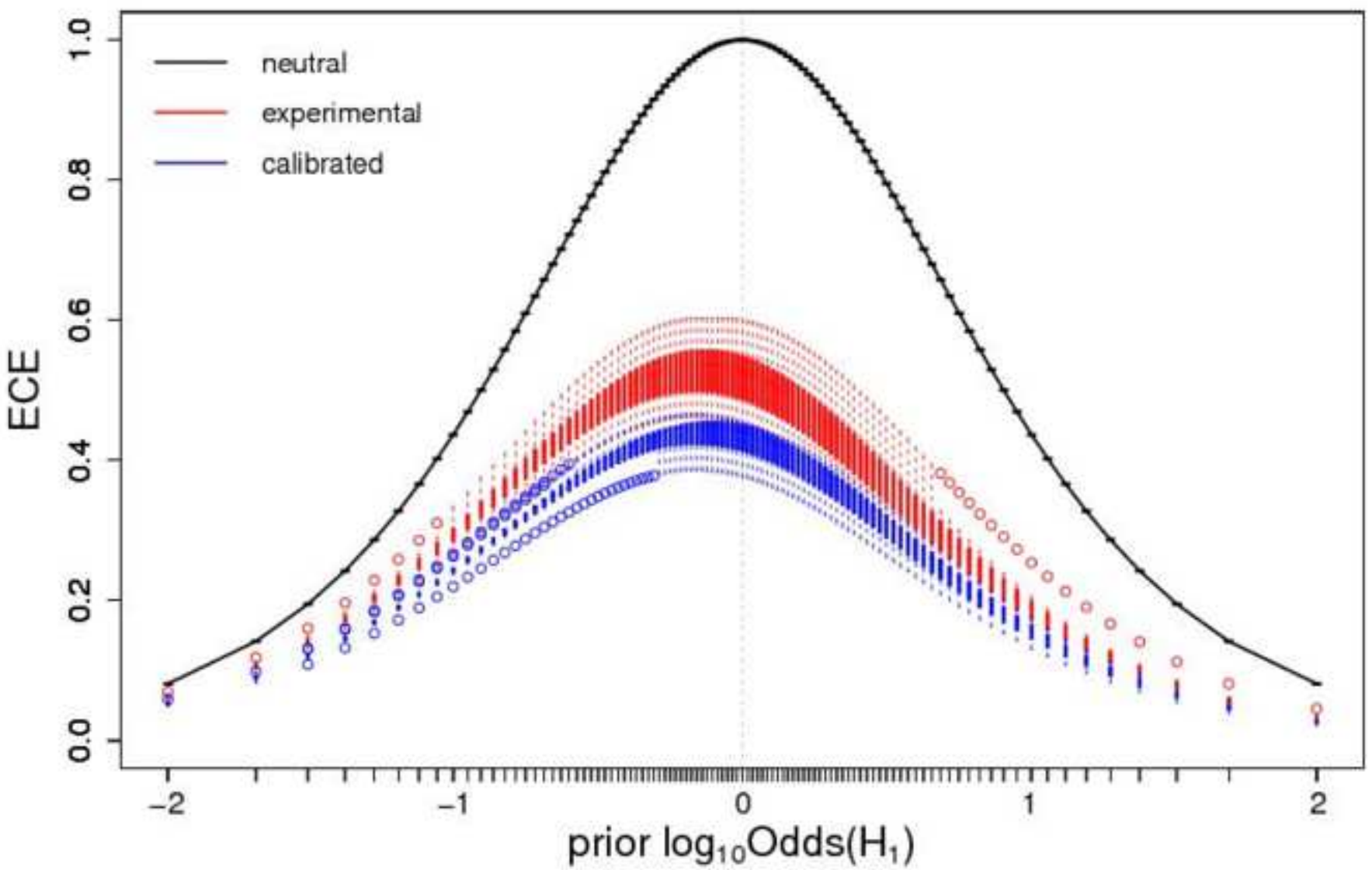


Figure 5d

Click here to access/download;Figure;Fig.5d.tif

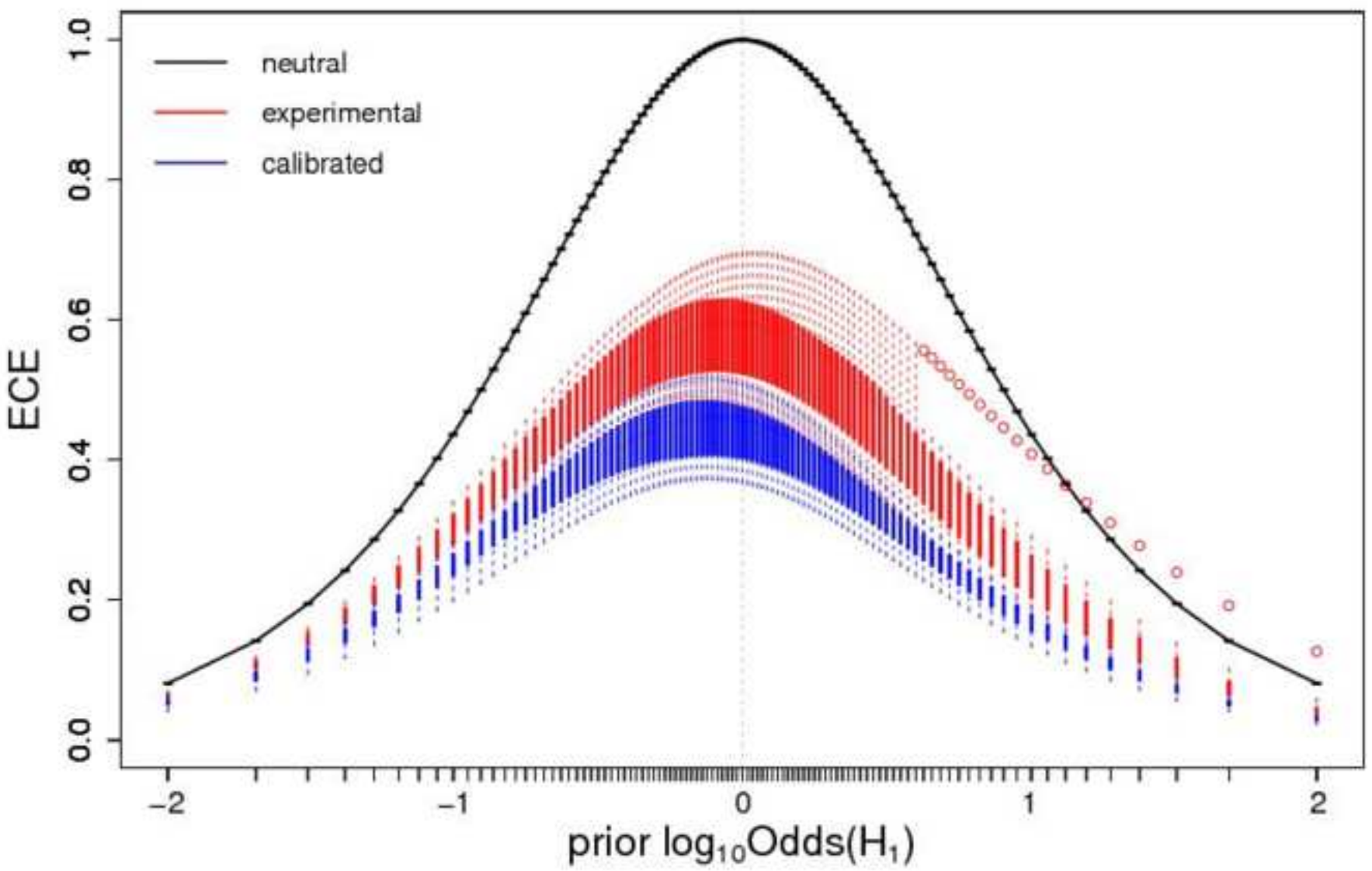


Figure 6a

Click here to access/download;Figure;Fig.6a.tif

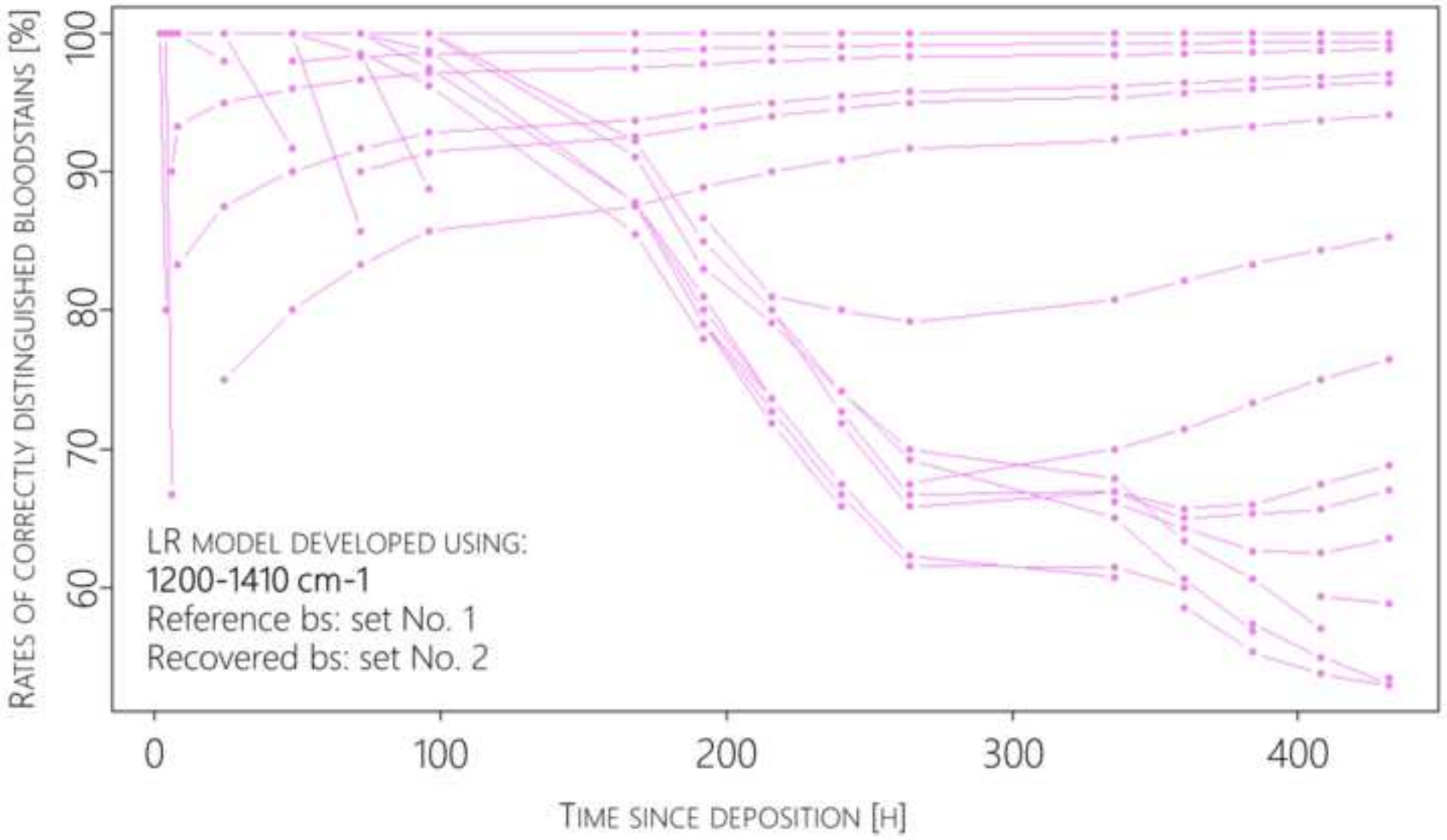


Figure 6b

Click here to access/download;Figure;Fig.6b.tif

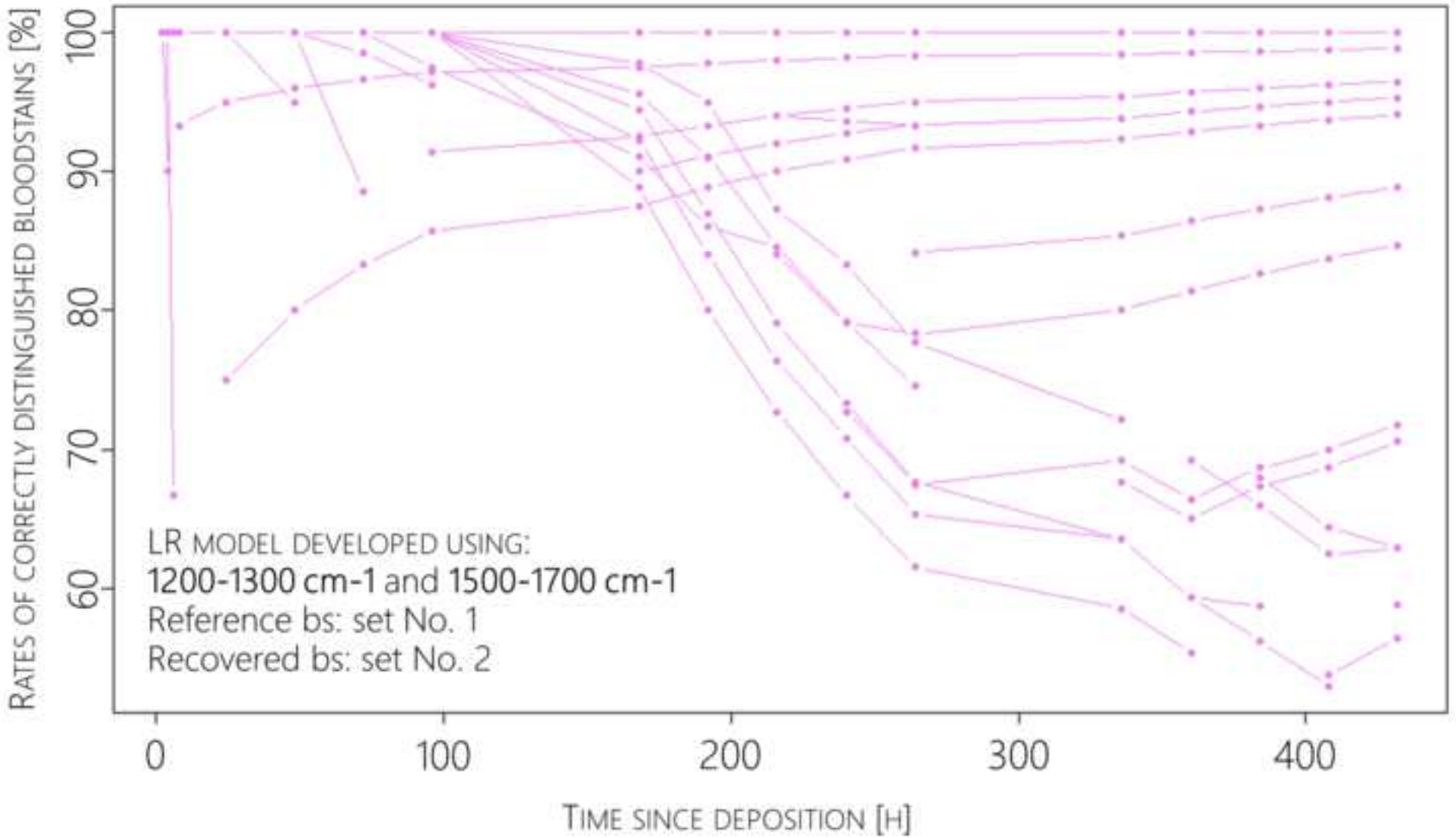
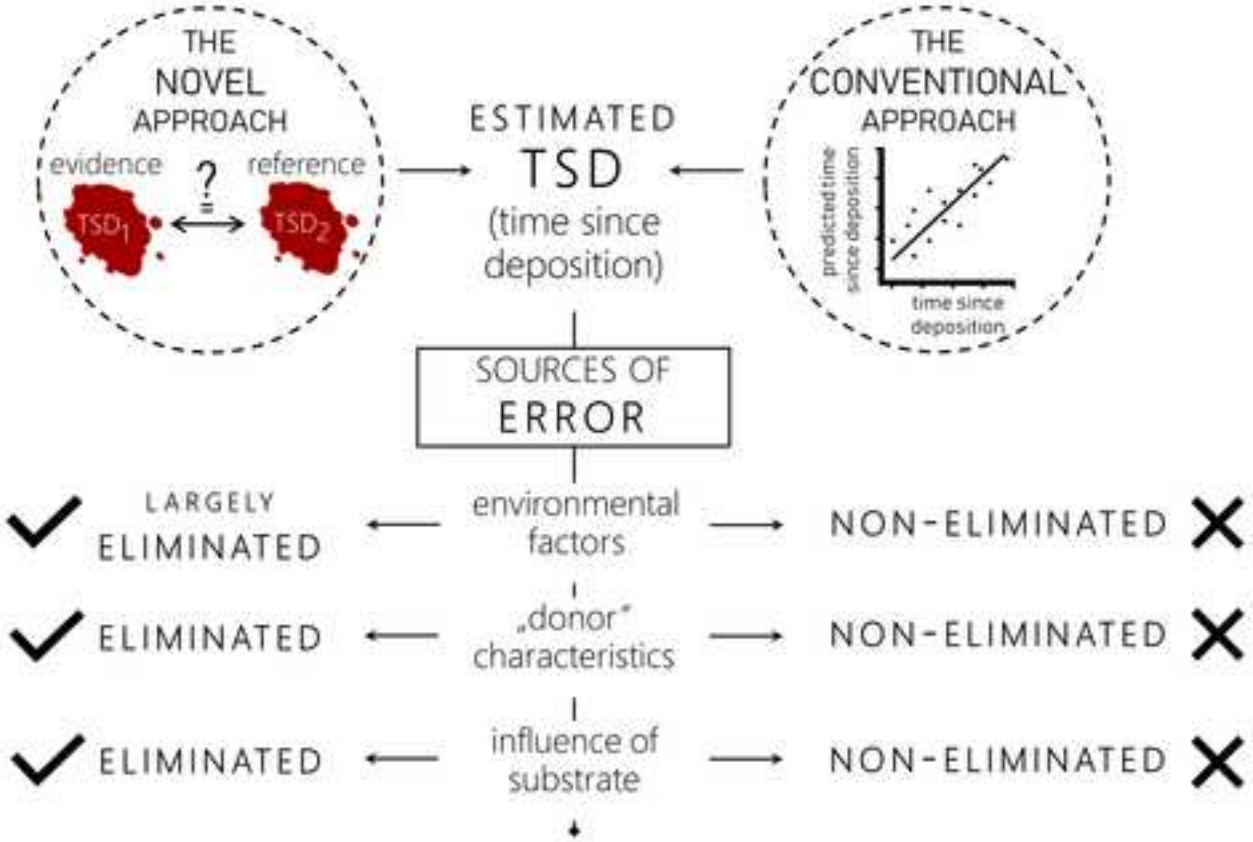


Figure 7 (revised)

Click here to access/download;Figure;Fig.7.tif



Response to the second review of the manuscript "**Breaking with trends in forensic dating: A likelihood ratio-based comparison approach**" with the reference number **FSI-D-23-00159R1**

We truly appreciate acknowledging the value of the new approach to the dating challenge for forensic purposes presented in the article. Most importantly, though, we are grateful for thoughtful feedback on the study. The constructive comments from the second Reviewer undoubtedly helped improve the manuscript. The responses to the Reviewer's remarks are provided below. They were also used to highlight certain important aspects of the proposed approach in the revised manuscript (all introduced changes have been highlighted in yellow in the article). We are also open to discussing any other comments and suggestions which may occur upon reading the responses to the review.

"The authors propose an interesting approach to a difficult problem where statistics should help a decision as accurate as possible from the scientific perspective. The probabilistic approximation may reduce the margin of error, but does not increase the accuracy in estimating the stain date, due to the importance of environmental factors and interactions with the characteristics of the support of the same (cotton, acrylic, glass, etc.) and the starting conditions of the same (stains, humidity, etc.). Only to the extent that the reference model spots approximate real conditions would it make practical sense."

Undoubtedly, no statistical method, no matter how advanced, will be sufficient to solve the problem of estimating the age of bloodstains. The same applies to the already and (most likely) subsequently proposed analytical methods presented as even more effective tools for monitoring blood degradation. Regardless of their level of advancement, none of them will provide the sought-after antidote if they are embedded in a conventional dating scheme aimed at developing a single universal model. Therefore, the mere application of probability-based models (such as the likelihood ratio method applied herein) would not demonstrate significant improvement either. The novelty of the approach presented in the manuscript does not lie in simply changing calibration models to probabilistic ones but rather in departing from the "dogma "of developing a single universal solution for such a diverse problem. It advocates for replacing it with a flexible approach to the dating task, tailored each time to the questioned evidentiary material and the conditions under which it was revealed. The essence of this solution has been presented in the first and the last sections of the manuscript entitled *1. Introduction* and *4. Conclusions and future perspectives*, respectively.

Therefore, throughout the article, it has been repeatedly emphasized that the most obvious but, at the same time, the most important conclusion of the presented research was that the high reproducibility of the supervised aging procedure is the determining factor for reliable and valid dating. It has been repeatedly noted in the *4. Conclusions and future perspectives* section, for example:

"(...) Another important note for future studies is that the high reproducibility of the supervised aging procedure is the determining factor for reliable and valid dating. (...)",

"(...) To reach valid TSD estimates, bloodstains serving as reference samples for building models must replicate the aging process of the evidence as accurately as possible. (...)",

"(...) Therefore, the success of the proposed approach will depend on how much the above factors influence degradation and, above all, on how much it is possible to control them during supervised aging. In this way, sources of error in TSD estimates that occur during the conventional dating approach would be largely eliminated (...)".

That being said, it can be stated that the approach proposed in the manuscript indeed is expected to "***increase the accuracy in estimating the stain date"*** (accuracy understood as the closeness of the TSD estimations to the "correct/real" value), as it is aimed at reducing the influence of external factors on the dating procedure. The influence of these external factors, such as environmental conditions, is often responsible for generating certain systematic errors that affect the accuracy of estimating TSD in the conventional approach, making it fall short of the ideal. For this reason, the authors acknowledge that:

"(...) future research efforts should identify factors that may affect the bloodstains' degradation – using experimental design methods – so that the supervised aging procedures focus on controlling only those specific factors".

Thus, we fully agree with the Reviewer's opinion that:

"The rigor of the expert evidence requires us to know the limitations of the model we apply and establish margins of error as accurate as possible. The work suggests a mathematical approach that would have an indicative value and that would require of a deeper critical analysis for its practical applications and to expose the limitations of the reference models that would require a design adjusted to the possible conditions of the problem spot. In summary, the value of a probability estimate would be directly related to the similarity of the conditions of the model to the problem spot."

and nowhere in the manuscript do we contradict that. We also do not attempt to present this new approach as a fully ready-to-use "product" for everyday forensic practice. While the solution proposed in the submitted manuscript has shown remarkable promise for more accurate TSD estimations, further work is still required before practical applications, as noted in one of the last paragraphs:

"Another important note for future studies is that the high reproducibility of the supervised aging procedure is the determining factor for reliable dating. In other words, a similar volume of blood, freshly drawn from the donor (if possible, the same individual who was the donor of the questioned trace), should be deposited on an identical substrate and stored under similar conditions as those observed at the crime scene. To reach valid TSD estimates, bloodstains serving as reference samples for building models must replicate the aging process of the evidence as accurately as possible. Irrespective of the monitoring method applied, the bloodstains' resultant aging profile depends on the aging kinetics, which is affected by such influential factors as the substrate, environmental conditions persisting during aging (controlled by climatic chambers), and the initial composition of the bloodstain. Therefore, the

success of the proposed approach will depend on how much the above factors influence degradation and, above all, on how much it is possible to control them during supervised aging. In this way, sources of error in TSD estimates that occur during the conventional dating approach would be largely eliminated (Fig. 7).

Obviously, it would be somewhat naïve to assume the possibility of recreating precisely the same conditions as evidence degradation. Nonetheless, recreating them as closely as possible to those prevailing during the degradation of the evidential bloodstain may be sufficient to answer the prosecutor's questions regarding the time of traces' formation. Thus, in the first place, future research should focus on identifying factors that may affect the bloodstains' degradation – using experimental design methods [51] – so that the supervised aging procedures focus on controlling only those specific factors. (...)"

As rightly pointed out by the second Reviewer, a deeper critical analysis of practical applications is necessary to expose the proposed procedure's limitations fully. This validation is currently being conducted using experimental design methods and environmental chambers. Details regarding this next research phase cannot be revealed at the current stage but will continue this publication cycle. However, in order to additionally highlight the importance of the proper validation of the proposed procedure, the following statement has been added:

"Indeed, the mere statement that errors arising from external factors will be largely eliminated is not a sufficient guarantee of the reliability of the proposed methodology. Hence, once again, it should be reminded that to meet the ISO/IEC 17025 international quality standards [16] or – simply – to ensure the validity of any novel methodology [51–53], a forensic laboratory must undertake a comprehensive evaluation of all factors that could impact the examination result during the validation process of a method intended for use in casework. In this specific case, after identifying the factors (and their combinations) that influence the degradation of bloodstains, the accuracy of TSD estimates using the new methodology can be fully assessed. It will be achieved by validating likelihood ratio models, including determining the rates of false positive and false negative responses, for "evidential" bloodstains degrading under different conditions, not only those presented in the present study. Only then will it be possible to consider advancing this methodology to practical application."

Finally, it should also be highlighted that this publication – as the title implies – aims to break a certain trend in forensic dating. At the same time, it is intended to be considered as an invitation addressed to other researchers to look at the problem from a different perspective, which may help solve it at least partially.

Once again, we thank the second Reviewer for the time spent reviewing our paper and look forward to meeting your expectations.

The Authors

CRediT authorship contribution statement

Alicja Menżyk: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Supervision.

Agnieszka Martyna: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - review & editing, Visualization.

Alessandro Damin: Methodology, Validation, Investigation, Resources, Writing - review & editing.

Marco Vincenti: Conceptualization, Methodology, Resources, Writing - review & editing, Supervision.

Grzegorz Zadora: Conceptualization, Formal analysis, Writing - review & editing, Supervision, Project administration.