Analytica Chimica Acta 1414 (2026) 345655

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Cross-study validation and guidance for resampling approaches for partial least squares discriminant analysis in spectral machine learning

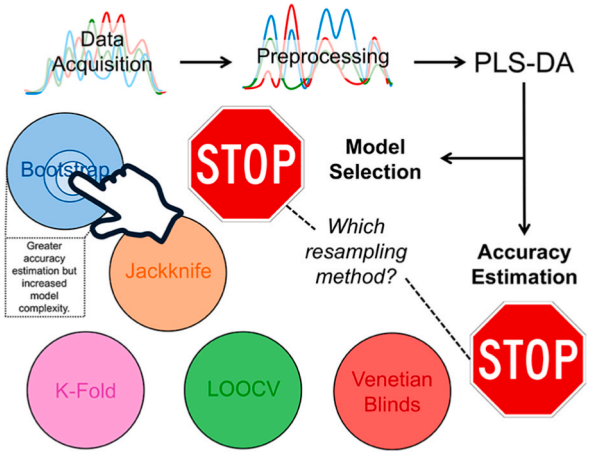
Aidan P. Holman [a,b], Dmitry Kurouski [a,b,*]

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

HIGHLIGHTS

- Spectral machine learning (SML) is a powerful tool that drives objective, high-throughput, and accurate identification across fields including medicine, food, and forensic analysis.
- SML outcomes depend on resampling strategies that can change model outcomes.
- Partial least squares-discriminant analysis (PLS-DA) models from nine studies are challenged with five different resampling methods.
- The results highlight that moderate, stable partitioning schemes offer the most robust and generalizable validation behavior for PLS-DA-based SML workflows.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Chemometrics
Machine learning
Resampling
Bootstrap
PLS

ABSTRACT

Resampling strategies are central to model selection and performance estimation in spectroscopic machine learning (SML), yet their impact is often treated as secondary to model choice. In this tutorial study, we systematically evaluated how common resampling methods influence model selection, predictive performance, and estimation bias within a fixed partial least squares-discriminant analysis (PLS-DA) framework across twelve independent spectroscopic datasets. Model tuning was performed using both the one-standard-error (1SE) rule and highest Macro F1 (HMF1) selection, and performance was assessed at the sample level using repeated external validation. Bootstrap and jackknife methods demonstrated the strongest ability to preserve predictive performance, whereas LOOCV and selected Venetian blinds and some K-fold approaches achieved a more favorable balance between performance stability and model parsimony. However, resampling strategies differed substantially in their ability to estimate true test performance. A Bayesian Bradley-Terry analysis revealed that K-fold cross-validation with five folds consistently produced the least biased estimates of external test Macro F1, outperforming both bootstrap and leave-one-out approaches. In contrast, LOOCV, jackknife, and Venetian blinds methods exhibited increased bias due to small or unrepresentative validation sets under sample-level resampling.

* Corresponding author. Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, United States.
E-mail address: dkurouski@tamu.edu (D. Kurouski).

https://doi.org/10.1016/j.aca.2026.345655
Received 9 January 2026; Received in revised form 20 March 2026; Accepted 12 May 2026
Available online 20 May 2026

These findings demonstrate that resampling strategy is a primary determinant of both model selection behavior and performance estimation accuracy in SML. This work provides a practical framework for selecting resampling methods that improve the reliability and interpretability of spectroscopic machine learning models.

1. Introduction

Spectral chemometrics refers to the application of mathematical and statistical methods to extract meaningful information from spectroscopic measurements. Within this domain, spectroscopic machine learning (SML) has emerged as a specialized branch that applies machine learning algorithms to spectral data, with particular emphasis on classification tasks [1–5]. In classification, the objectives can be broadly defined as: (1) assigning unknown spectra to predefined categories with maximal accuracy, (2) ensuring that models generalize reliably to new, independent data, and (3) interpreting underlying spectral features in a way that supports scientific or applied decision-making [6].

One common strategy for evaluating and optimizing classification models in SML is the use of resampling techniques [7]. Resampling involves repeatedly partitioning or drawing from the available dataset to estimate how well a model is expected to perform on unseen data. Approaches such as leave-one-out cross-validation (LOOCV), K-fold cross-validation (KFCV), Venetian blinds partitioning (VBCV), bootstrapping (BS), and jack-knifing (JK) are frequently employed to guide hyperparameter tuning and performance estimation [5,8]. While these approaches are theoretically intended for different utilizations—cross-validation (CV) being designed for estimating external test performance and hyperparameter selection, and sampling-with-replacement procedures like BS and JK for error estimation and assessing model stability—they can nonetheless be applied toward similar objectives of model evaluation and optimization through validation of the given splits or folds [9]. Additionally, their accuracy and reliability vary depending on data size, structure, and algorithmic choice [7].

One widely used SML model is partial least squares-discriminant analysis, or PLS-DA, which has been adopted across diverse fields, including food and agricultural authentication, biomedical diagnostics, environmental monitoring, and forensic analysis, where robust classification is often critical. For instance, near-infrared spectroscopy combined with PLS-DA has been used to authenticate olive oil and detect adulteration in milk powders [10–12]; Raman spectroscopy has been applied to differentiate between malignant and benign tissue in biomedical diagnostics [13–15]; UV-Vis and fluorescence spectra paired with classification algorithms have enabled monitoring of water pollutants and heavy metal contamination [16–18]; and infrared and Raman-based models have been developed to identify synthetic dyes in forensic trace evidence such as textiles and hair [19–23]. Despite this breadth of application, resampling and testing strategies remain inconsistent, not only across different laboratories but even within single research groups. We are not suggesting that the scientific validity of those studies is in question; rather, differences in SML tuning practice may introduce variability in reported performance, and stabilizing our reliance on more consistent resampling methods could improve reproducibility and comparability. Furthermore, such variability complicates comparisons across studies and raises questions about whether certain resampling schemes systematically provide more reliable predictions than others.

In this work, we re-examined previously published PLS-DA models across multiple studies. Specifically, we address two questions: (i) does re-optimizing model latent variables under alternative resampling schemes lead to improved test performance, and (ii) which resampling strategies consistently approximate external test prediction metrics across different studies? By systematically comparing strategies such as LOOCV, KFCV, VBCV, BS, and JK, we aim to clarify the strengths and limitations of these approaches and provide practical guidance for their use in future SML research involving PLS-DA.

Table 1
Characteristics of the twelve studies included in this work. All manuscripts and datasets are referenced after the relevant study.

Research Group	Study	Type of Specimen	Type of Spectroscopy	Classification Objective	No. of Classes	No. of Independent Samples per Class	Total No. of Spectra
Internal	Goff et al. (2022) [24,25]	(Cannabis) plant	Raman	Sex of plant.	3	22-24	213
	Higgins et al. (2022) [26,27]	(Wheat) plant	Raman	Type of stress on plant.	5	29-30	231
	Holman and Kurouski (2023) [22,28]	Dyed hair	Raman	Specific dye on hair.	4	10	2000
	Holman et al. (2024) [29,30]	(Secondary Screwworm) fly larva	Infrared	Sex of larva.	2	11-12	115
	Rodriguez et al. (2025) [31,32]	Plant	Raman	Lunar or earth-grown.	2	15-16	59
	Holman et al. (2025a) [33,34]	(Hairy Maggot Blowfly) fly larva	Infrared	Sex of larva.	2	31-34	325
	Juárez et al. (2025) [35,36]	(Human) blood	Raman	Infected with *Borrelia burgdorferi*.	2	98-99	1291
	Sasaki et al. (2026) [37,38]	Cotton Fabric	Raman	Dye on fabric across exposure periods.	2	16	739
External	Kosmowski and Worku (2018) [39]	(Sorghum) kernels	Infrared	Cultivar of Sorghum kernels.	10	50	500
	Muthreich et al. (2020) [40,41]	Pollen	Infrared	*Quercus* spp. of pollen.	6	15-75	920
	Banerjee et al. (2021) [42,43]	(Human) blood plasma	Infrared	Severity of COVID-19.	2	52-78	260
	Barney et al. (2025) [44,45]	(Human) nails	Infrared	Fentanyl exposure.	2	16-63	1185

a: Before 50/50 validation-set approach was applied.

2. Materials and methods

2.1. Study selection

Twelve previously published studies within (8) and outside (4) our laboratory that include PLS-DA modelling were considered for analysis (Table 1). To ensure sufficient statistical robustness, studies were selected only if they contained spectral datasets with a minimum of ten independent samples (not spectra) per class, thereby reducing the risk of artificially inflated performance from under-sampled categories. The diversity of specimens (plants, insects, biological fluids, and forensic materials), spectroscopic platforms (Raman and infrared), and classification objectives (sex determination, pathogen infection, stress response, and material identification) provides a representative cross-section of the application space for spectral machine learning.

2.2. Data analysis

All chemometrics and plotting were performed using Python (v13.3). However, some preprocessing was necessarily done in MATLAB using PLS_Toolbox 9.5.

2.2.1. Preprocessing

All data was preprocessed exactly as described in their given study, before partitioning or resampling, and detailed in Table S1. However, to avoid data leakage during model training and thereby increase robustness of results, mean and median-centering-based (as well as similar) processing methods were excluded from study datasets where relevant (Table S1).

2.2.2. Machine learning

All of the PLS-DA modelling, Python libraries, and parameter tuning strategies are summarized in Table S2. The specific range of latent variables (LVs) chosen for automatic tuning (LVs 2-20) is because beyond 20, added components mostly model noise or idiosyncratic structures and a 1 LV model rarely captures sufficient class-relevant covariance [46]. The one exception is the dataset for Kosmowski and Worku (2018) which was initially trained across 50 LVs, with 44 LVs as the optimal model. If we were to only subject the model to up to 20 LVs, parsimony (selected model with lower LVs than baseline) would be achieved in all cases. Therefore, this dataset must be compared to the same range of initial LVs in order to draw conclusions on the relative performance of resampling methods.

In this study, multi-class PLS-DA was implemented using the PLSRegression algorithm in scikit-learn, which supports multivariate response modeling (sometimes called PLS2). Class membership was encoded using a dummy (binary indicator) response matrix, allowing all classes to be modeled simultaneously within a single PLS framework [47]. This formulation does not impose any ordinal structure on the class labels and is consistent with the standard definition of multi-class PLS-DA described in the chemometric literature.

Parameter selection was performed using the one-standard-error (1SE) heuristic approach and the highest score approach (HMF1, for highest Macro F1) from CV-derived Macro F1 scores, two widely recommended model-selection rules in chemometrics and cross-validation literature [48]. The 1SE rule selects the simplest model (i.e., the lowest number of LVs) whose validation performance lies within one standard error of the maximum. This approach is especially appropriate for PLS-DA, where the primary tuning parameter is the number of LVs, and overestimation of LVs is a known source of overfitting in spectral classification. Under this rule: (i) if CV accuracy peaks at LV = 15 but LVs = 7-14 fall within one standard error of the peak, the model selects LV = 7; (ii) if accuracy increases monotonically until LV = 10 and then declines, LV = 10 is chosen; and (iii) when the CV curve oscillates, the first statistically indistinguishable local maximum is selected rather than the global maximum. This promotes parsimony, reduces variance, and avoids selecting LVs that model noise rather than chemically meaningful variance.

On the other hand, the HMF1 rule selects the model that achieves the highest mean validation Macro F1 score across validation splits/folds and represents the most widely used model-selection strategy in machine learning [49]. This approach directly targets predictive performance and is commonly used for tuning classification models when the primary objective is maximizing generalization accuracy. In chemometrics and broader statistical learning literature, selecting the model with the best cross-validated performance (e.g., minimum CV error or maximum accuracy) is a standard alternative to parsimony-based heuristics such as the 1SE rule [49]. By evaluating both approaches, the present study aims to capture the range of models typically selected in practice, from parsimonious solutions favored by the 1SE rule to performance-maximizing solutions obtained through the HMF1 criterion.

Other selection procedures, such as total efficiency (TEFF) under soft probabilistic discrimination of ranked classes, are designed to optimize classification performance by jointly weighting sensitivity and specificity across classes and finding the parameter where validation and testing agree on performance estimation [50]. In contrast, the 1SE rule does not aim to maximize a performance metric but instead controls model complexity by explicitly accounting for the uncertainty in the cross-validation estimate. This distinction is critical in PLS-DA, where increases in apparent discrimination with additional latent variables frequently fall within the noise of the validation process and do not translate to improved generalization. Additionally, since the TEFF approach requires the peaking of test metrics to select parameters, we are unable to conduct unbiased baseline test comparisons.

2.2.3. Resampling strategies

All resampling strategies (Table S3) were implemented using scikit-learn (sklearn), with additional control of class imbalance through *RandomOverSampler* from imblearn, applied only to the training partition within each resampling iteration to ensure the test split remained representative of the original distribution. The choice of fold numbers (K) and split parameters (S for Venetian blinds and Jackknife, B for Bootstrap) was guided by values most commonly reported in the statistical learning and chemometric validation literature, ensuring alignment with established practice rather than arbitrary selection [48,51,52]. Where applicable, StratifiedKFold was used to enforce balanced class representation across folds (Table S3). For LOOCV, for example, stratification was not enforced because only one sample is held out per split, meaning only a single class is represented in the test set, rendering stratification ineffective.

Model predictive performance was measured using the Macro F1 score at the sample level via f1_score from sklearn. The F1 score balances precision and recall and gives a greater idea of accuracy in an unbalanced-sample model, calculated as the number of true positives (TPs) divided by the number of TPs plus one-half of the total false positives and false negatives.

2.2.4. Bayesian estimation

A Bayesian framework was used to evaluate whether each resampling method was more likely to help than harm model performance across studies. A Bayesian approach is more appropriate than a frequentist alternative for this setting because (i) the number of studies per method is relatively small (n = 12), making frequentist p-values unstable and sensitive to single outcomes; (ii) Bayesian posterior distributions naturally incorporate uncertainty and avoid the dichotomous interpretation inherent in frequentist hypothesis testing; and (iii) hierarchical or method-level variation is better represented by probability distributions than by point null hypotheses. Importantly, the Bayesian approach allows direct interpretation of quantities such as P (method improves accuracy or it worsens), which aligns with how practitioners actually make decisions about resampling strategies.

For example, consider a hypothetical evaluation of a resampling strategy applied across twelve independent spectroscopic studies similar to those examined here. Suppose the method improved validation accuracy in seven studies, showed no change in two, and slightly reduced accuracy in three. A frequentist analysis might test whether the resampling strategy improved or did not change accuracy more than it diminished. In this case, using a one-sided sign test for example, you would need at least 10 out of 12 studies to have showed improved or unchanged accuracy to be considered significant (one-sided sign test for 9/12 studies' $p = 0.073$). On the other hand, a Bayesian Beta-Binomial model with Jeffreys prior yields a posterior distribution of Beta(9.5, 3.5) for the probability that the method is non-harmful. From this posterior, the probability that the method does not harm accuracy (i.e., $p > 0.5$) is approximately 0.94, and the posterior mean probability of a non-harmful outcome is ~0.73. Thus, even though the frequentist test does not reach conventional significance, the Bayesian analysis indicates strong evidence that the method is more likely to help (or at least not harm) than to degrade performance, providing a more informative and decision-relevant interpretation under limited sample sizes.

The canonical Bayes' theorem states that:

$$p(\theta|y) = \frac{p(y|\theta)^{*}p(\theta)}{p(y)},$$

Where $p(\theta|y)$ is the posterior probability that event θ occurred given the evidence y, $p(y|\theta)$ is the likelihood of evidence occurring due to the event θ, $p(\theta)$ is the prior information, and $p(y)$ is the marginal likelihood.

Jeffrey's prior [Beta(0.5,0.5)] was considered since it has higher uncertainty (i.e., it is more conservative) for smaller sample sizes compared to a uniform prior. A Beta-Binomial model was used to compare resampling strategies across studies. In this framework, the number of non-harmful outcomes (e.g., improvements or no change in performance) is modeled as a Binomial process with an unknown probability parameter, while the Beta distribution serves as a conjugate prior over this probability. This yields a closed-form posterior distribution that directly quantifies the probability that a given resampling method is beneficial across studies. The formula can be written as:

$$y \sim Binomial(n, p),$$

Let y be the number of "non-harmful" outcomes out of n studies and $p = \theta$, for the probability of non-harmful outcomes. Using Jeffrey's prior and substituting for Bayes' theorem:

$$p|y \sim Beta(y + 0.5, n - y + 0.5),$$

and:

$$P(p > 0.5|y) = 1 - F_{Beta}(0.5;\ y + 0.5, n - y + 0.5),$$

where F_{Beta} denotes the cumulative distribution function of the Beta distribution. Equivalently, this probability can be expressed as:

$$P(p > 0.5|y) = \int_{0.5}^{1} Beta(p; y + 0.5, n - y + 0.5)dp.$$

A Bayesian Bradley-Terry model was used to compare resampling strategies based on their squared bias in estimating test Macro F1. Rather than performing explicit pairwise comparisons within individual studies, all observations for each method were compared against those of every other method, and the results were aggregated at the method-pair level. Specifically, for each unordered pair of methods i and j, the total number of non-tied comparisons n_{ij} and the number of times method i produced a lower squared bias than method j, denoted w_{ij}, were computed. These aggregated outcomes were modeled as:

$$w_{ij} \sim Binomial\left(n_{ij}, \pi_{ij}\right),$$

where π_{ij} represents the probability that method i outperformed method j. The Bradley-Terry model defines this probability through a logit link that is equivalent to:

$$\pi_{ij} = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}},$$

where θ_i and θ_i are latent skill parameters representing the relative performance of each method. To ensure identifiability of the model, the skill parameters were constrained such that:

$$\sum_{k=1}^{K} \theta_k = 0,$$

So that each parameter represents performance relative to the average method.

The model was implemented in the PyMC probabilistic programming framework, and posterior summaries and diagnostics were generated using ArviZ. Inference was performed using the No-U-Turn Sampler (NUTS) with four independent Markov Chain Monte Carlo (MCMC) chains, each consisting of 1000 tuning iterations followed by 1000 posterior draws, for a total of 4000 samples after warm-up. These settings represent established defaults for high-quality Bayesian estimation and were adequate to achieve convergence and effective sample sizes appropriate for downstream inference [53].

2.2.5. Tests of significance

Although Bayesian inference was the primary framework used to quantify uncertainty and support decision-making, frequentist hypothesis tests were employed as complementary diagnostic tools to guide and contextualize the Bayesian analysis. Specifically, nonparametric, Wilcoxon signed-rank tests were used to assess the direction and consistency of paired differences prior to probabilistic modeling, ensuring that observed effects were not driven by outliers or systematic violations of assumptions. This dual-framework approach provides coherence with established analytical practices in chemometrics and machine learning while allowing Bayesian models to build upon empirically validated patterns in the data rather than relying solely on prior assumptions or model structure.

Nonparametric, Wilcoxon signed-rank tests were performed throughout the study for the following reasons: (1) the Wilcoxon signed-rank test is specifically designed for paired comparisons and is robust to non-normal, skewed, or heteroscedastic data commonly observed in resampling-based machine learning evaluations [48,54,55]; and (2) it has been widely recommended in machine learning and chemometric benchmarking studies as a more appropriate alternative to the paired *t*-test when comparing model performance across repeated splits or folds [54,55].

3. Results and discussion

In statistical learning theory, the No Free Lunch (NFL) theorem asserts that no single model or algorithm outperforms all others across every possible dataset or problem domain; performance is inherently data-dependent [56]. Therefore, the purpose of this study is not to identify a universally optimal SML model but to evaluate how different resampling strategies affect the performance within the same model architecture under identical data conditions. By holding the model constant and only varying the resampling method, valid comparative inferences can be made regarding which strategy yields better generalization behavior for that specific model-data pairing.

Furthermore, if a particular resampling approach demonstrates consistently strong performance across multiple datasets evaluated using the same model, it can be reasonably generalized as a robust strategy for future projects involving that model class; while still acknowledging that this conclusion is conditional on similarity in data structure and distribution, not an absolute optimization claim. This interpretation aligns with best practices in model validation and empirical risk minimization, where consistency across resampling

Table 2
Initial (baseline) test performance of each study's model and dataset using 10 random 50-50 train-test partitions.

Study	Baseline Resampling Strategy	Baseline Parameter[a]	Test Median Macro F1 Score, % (SE)	Supporting Tables
Goff et al. (2022)	Unclear	LVs = 16	75.34 (2.39)	Tables S4 and S16
Higgins et al. (2022)	Unclear	LVs = 8	95.99 (0.61)	Tables S5 and S17
Holman and Kurouski (2023)	Unclear	LVs = 3	100.0	Tables S6 and S18
Holman et al. (2024)	Unclear	LVs = 11[b]	82.86 (5.28)	Tables S7 and S19
Rodriguez et al. (2025)	Unclear	LVs = 6[b]	93.73 (2.20)	Tables S8 and S20
Holman et al. (2025a)	KFCV (K = 10)	LVs = 8	85.76 (1.91)	Tables S9 and S21
Juárez et al. (2025)	Unclear	LVs = 13	85.01 (1.05)	Tables S10 and S22
Sasaki et al. (2026)	VBCV (S = 10)	LVs = 6	87.50 (2.17)	Tables S11 and S23
Kosmowski and Worku (2018)	KFCV (K = 5)	LVs = 44	83.33 (0.57)	Tables S12 and S24
Muthreich et al. (2020)	KFCV (K = 100)	LVs = 4	62.58 (1.14)	Tables S13 and S25
Banerjee et al. (2021)	LOOCV	LVs = 4	71.23 (1.42)	Tables S14 and S26
Barney et al. (2025)	Unclear	LVs = 8	75.92 (1.17)	Tables S15 and S27

[a] All baseline parameters are based on reported or verified selections in the relevant manuscript; **b**: These studies reported one less latent variable in their final model, but upon re-examination of the data and models, the correct ones are now listed.

evaluations is taken as evidence of stability and practical utility rather than universal superiority [48,57].

3.1. Re-tuning model parameters using 1SE rule approach

To determine which resampling strategies are efficient for automatic tuning across different studies, we compared resampling strategies on the same data, model, and LVs by estimating metrics at the sample-level. To determine our baseline test performance (i.e., the fixed standard model metrics for comparison), we first estimated expected test Macro F1 for each study's model and dataset (Table 2) by intentionally using a ~50/50 hold-out (50% for training/tuning; the remainder for testing) randomly split 10 times to stress-test the resampling strategies under limited training data while maximizing the size of the independent test set. This approach is called the validation set approach and typical studies employ 60/40, 70/30, or 80/20 splits for training and testing their models, respectively [57–59]. Our specific partition is designed to decrease the variance in the test accuracy estimate (Macro F1 in our case) compared to typical splits [60]. The 10 random splitting events allow us to incorporate uncertainty into our median estimates of Macro F1 scores and perform tests of significance in order to conclude if one resampling method performed better, worse, or no different than the original study's LV-selected model.

After running LV tuning for each resampling method across each study, we found that resampling strategies exhibit markedly different behaviors in both performance preservation and model complexity, Fig. 1 and Tables S4–S15. Bootstrap and jackknife approaches (BS and JK) show the lowest proportions of Macro F1 decreases, with the majority of studies falling into the "No Change" or "Improved" categories, indicating strong stability in predictive performance. In contrast, KFCV (K = 5) exhibits the highest rate of Macro F1 decreases, suggesting greater risk of performance degradation. With respect to latent-variable selection, LOOCV and Venetian blinds cross-validation (particularly S = 5 and S = 10) display comparatively higher proportions of reduced latent variables while still maintaining favorable performance profiles, reflecting a more balanced trade-off between stability and simplicity. Overall, bootstrap and jackknife prioritize performance stability, whereas LOOCV and selected Venetian blinds configurations achieve a more balanced compromise between preserving accuracy and encouraging parsimonious model selection.

Interestingly, differences in resampling performance between model selection strategies (1SE vs. HMF1) revealed a clear trade-off (Fig. 1). The HMF1 approach consistently resulted in lower degradation of Macro F1, indicating improved predictive performance, but this came at the expense of reduced model parsimony (i.e., selection of more complex models). In contrast, both jackknife (JK) approaches exhibited slightly greater Macro F1 degradation under HMF1 compared to the 1SE criterion. Overall, these findings suggest that while HMF1 is more likely to yield higher-performing models, it does so by favoring increased model complexity.

To formally quantify the reliability of each resampling method, we implemented a Beta-Binomial Bayesian framework. Rather than simply reporting the proportion of studies in which performance was preserved, this approach allowed us to estimate the underlying probability that a given resampling strategy would yield a non-worse Macro F1 model selection, along with the conditional probability of parsimony given that performance was not degraded (Fig. 2). Under this approach, we obtained posterior means, credible intervals, and posterior probabilities that each event exceeded chance (50%), thereby quantifying both effect size and strength of evidence, Table 3.

Under the Bayesian Beta–Binomial framework, clear differences emerged in the stability of Macro F1 model selection across resampling strategies (Table 3). Bootstrap and jackknife methods demonstrated the strongest evidence for preserving performance. Specifically, BS (B = 100) and BS (B = 1000) yielded posterior mean probabilities of non-worse Macro F1 of 90.0% (95% CrI: 75.6-98.2) and 94.0% (95% CrI: 82.1-99.5), respectively, with posterior probabilities exceeding 99.9%, indicating near certainty that these methods outperform chance in maintaining Macro F1. Similarly, JK (S = 100) and JK (S = 1000) achieved posterior means of 86.0% (95% CrI: 70.3-96.4) and 90.0% (95% CrI: 75.9-98.2), respectively, again with posterior probabilities >99.9%, reflecting very strong evidence of performance stability. LOOCV and several Venetian blinds configurations also performed favorably; for example, LOOCV and VBCV (S = 10) both demonstrated posterior means of 78.0% (95% CrI: 60.2-91.6) with posterior probabilities of 99.8%, while VBCV (S = 5) yielded a mean of 74.0% (95% CrI: 55.5-88.8) with posterior probability of 99.4%, indicating robust (though slightly less extreme) evidence for non-worse Macro F1 selection.

In contrast, not all k-fold configurations exhibited equivalent stability. While KFCV (K = 3) showed strong performance (78.0%, 95% CrI: 60.2-91.6; P > 0.5 = 99.8%), KFCV (K = 10) was based on fewer studies (n = 14) but still demonstrated high probability of non-worse performance (76.7%, 95% CrI: 53.1-93.6; P > 0.5 = 98.6%). However, KFCV (K = 5) showed notably weaker evidence, with a posterior mean of 62.0% (95% CrI: 42.6-79.6) and posterior probability of 89.0%. Although still above chance, this reduction suggests greater variability and less reliable performance preservation relative to other configurations. These results indicate that k-fold cross-validation should not be treated as a single homogeneous method, as specific fold choices materially influence model-selection reliability.

When parsimony was examined conditionally among studies in which performance did not worsen, a second pattern emerged that revealed an important trade-off. LOOCV and KFCV (K = 3) demonstrated strong evidence for parsimony, with posterior mean probabilities of 72.5% (95% CrI: 51.6-89.2) and 77.5% (95% CrI: 57.4-92.4), respectively, and posterior probabilities of 98.2% and 99.5%. VBCV (S = 3) and VBCV (S = 5) also showed favorable balance, with means of 69.4% (95% CrI: 47.0-87.8; P = 95.7%) and 65.8% (95% CrI: 43.7-84.7;

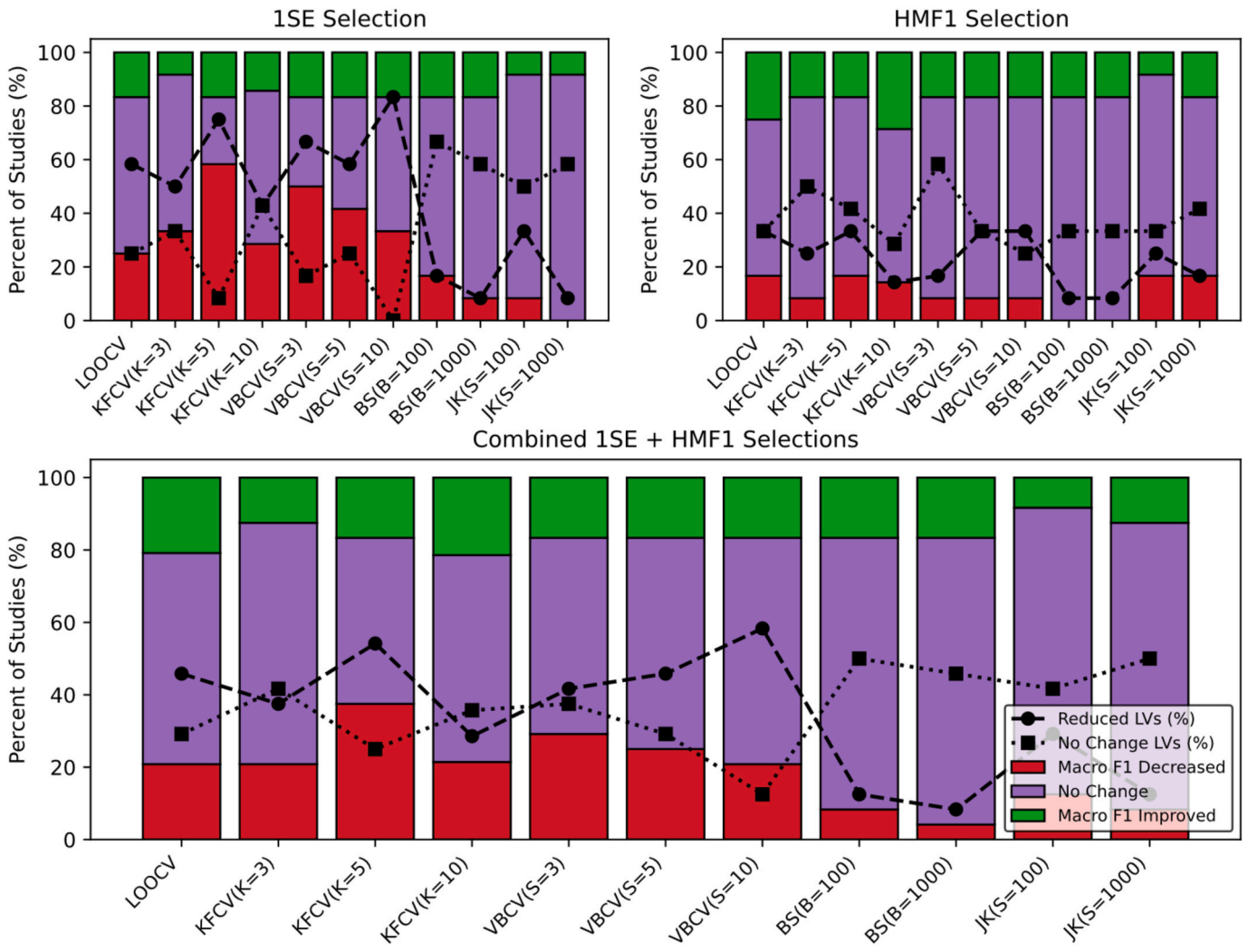


Fig. 1. Performance characteristics of each resampling strategy across all datasets. Stacked bars represent the distribution of Macro F1 outcomes relative to the baseline test model: decreased Macro F1 (red), no change (purple), or improved Macro F1 (green). The lines show parsimony outcomes, including the percentage of studies with reduced latent variables (solid circles) and no change in latent variables (solid squares). *It should be noted that KFCV (K=10) was only applied to 7 out of 12 studies due to some possessing at least one class having less than 10 independent samples during training after partitioning.* (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

P = 92.2%), respectively, indicating consistent support for selecting simpler models when Macro F1 was preserved. In contrast, KFCV (K = 10) showed only weak evidence for parsimony (54.2%, 95% CrI: 27.0-80.0; P = 61.8%), suggesting inconsistent complexity control.

By contrast, bootstrap and jackknife methods, despite their near-certain preservation of Macro F1, showed weaker and more variable evidence for parsimony. For example, BS (B = 100) and BS (B = 1000) yielded conditional parsimony means of 58.7% (95% CrI: 38.5-77.5; P = 80.3%) and 52.1% (95% CrI: 32.5-71.3; P = 58.2%), respectively, indicating only modest support for selecting simpler models. Similarly, JK (S = 100) and JK (S = 1000) produced means of 65.9% (95% CrI: 45.4-83.7; P = 93.8%) and 58.7% (95% CrI: 38.5-77.5; P = 80.3%), respectively. These findings indicate that although these methods are highly stable in preserving predictive performance, they are comparatively less consistent in promoting model parsimony, suggesting a tendency toward more complex model selection.

Taken together, the results indicate that resampling strategies differ primarily in their trade-off between performance stability and parsimony. Bootstrap and jackknife approaches maximize stability of Macro F1 but provide weaker and less consistent evidence for complexity reduction. LOOCV, KFCV (K = 3), and VBCV (S = 3-5) demonstrate strong evidence for both non-worse performance and parsimony, representing more balanced model-selection strategies. In contrast, KFCV (K = 5) exhibits comparatively weaker stability and parsimony support and may therefore be less reliable in similar analytical contexts.

3.2. *Efficient resampling methods for estimating test performance*

As stated earlier, strategies in SML research remain highly heterogeneous, and some studies report only cross-validation-based accuracy estimates without performing external validation on an independent hold-out dataset. Selecting LVs = 10 because K-fold cross-validation (KFCV) minimizes the calibration misclassification rate does not imply that the cross-validated accuracy at that point is an unbiased estimate of the model's final test accuracy. In other words, while cross-validation is useful for internal model comparison, it does not replicate the statistical independence of a true external test set. When hyperparameters are tuned using cross-validation and the same cross-validation estimates are then reported as "test accuracy," performance estimates may be inadvertently inflated due to information leakage or selection-induced bias. External validation, by contrast, evaluates the fully specified model on data that played no role in model tuning, thereby providing a more defensible estimate of real-world predictive performance. For these practitioners an important question arises: which resampling methods

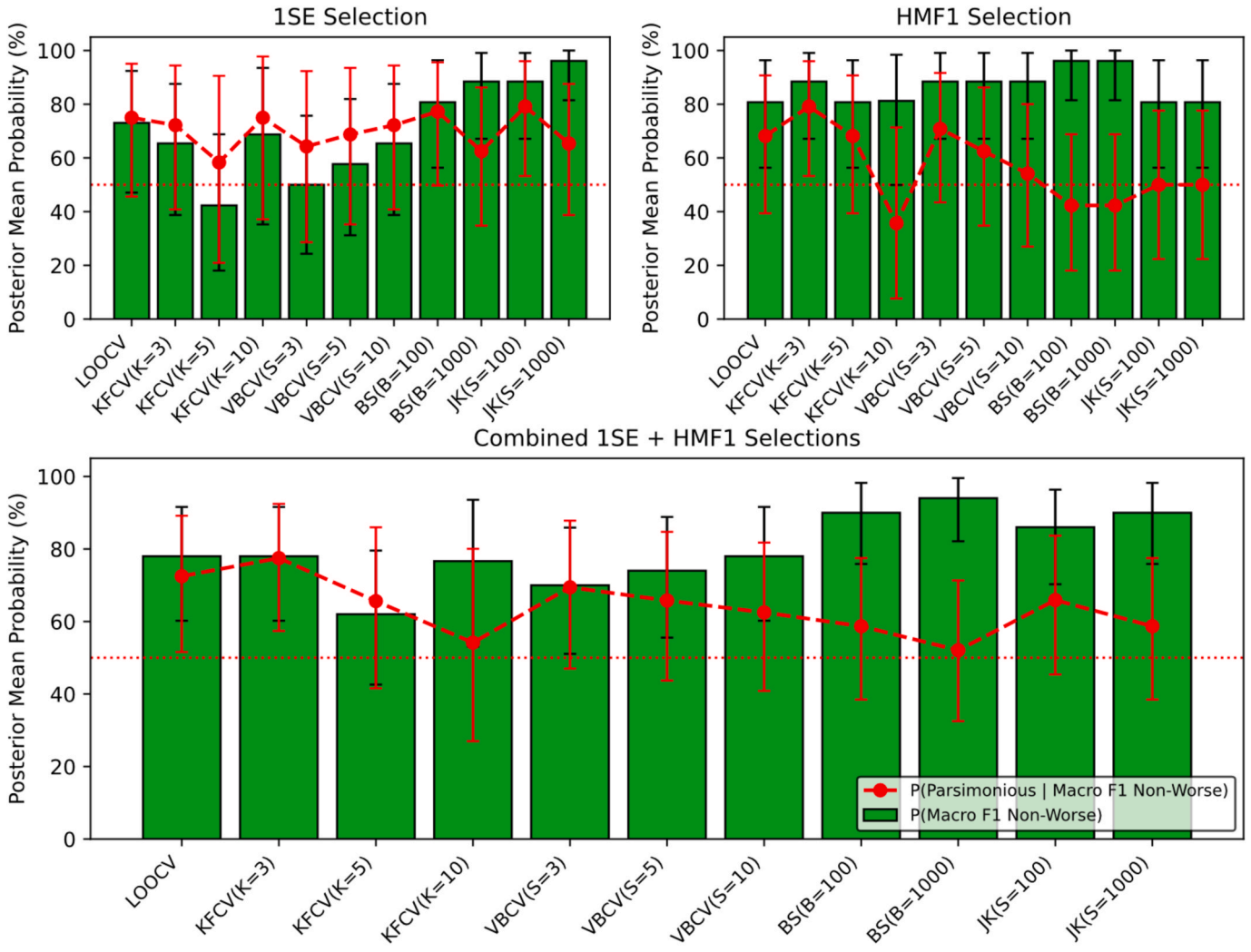


Fig. 2. Posterior mean probability that each resampling method yields non-worse Macro F1 (green bars) and the mean conditional probability that the method is parsimonious given non-worse Macro F1 (black dashed line). The red dotted line indicates a reference probability of 0.50. *It should be noted that KFCV (K=10) was only applied to 7 out of 12 studies due to some possessing at least one class having less than 10 independent samples during training after partitioning.* (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Summary results for combined 1SE and HMF1 selections' estimated probability of resampling method's non-worse Macro F1 model selection (A) and parsimony (B) given non-worse Macro F1 model selection. Additionally, the posterior probabilities of A and B events occurring beyond chance (50%).

Method	N studies	Mean P(A) = p_1, % (95% CrI)	P ($p_1 > 0.5$), %	Mean P(B|A) = p_2, % (95% CrI)	P ($p_2 > 0.5$), %
LOOCV	24	78.0 (60.2, 91.6)	99.8	72.5 (51.6, 89.2)	98.2
KFCV (K = 3)	24	78.0 (60.2, 91.6)	99.8	77.5 (57.4, 92.4)	99.5
KFCV (K = 5)	24	62.0 (42.6, 79.6)	89.0	65.6 (41.6, 86.0)	90.2
KFCV (K = 10)	14	76.7 (53.1, 93.6)	98.6	54.2 (27.0, 80.0)	61.8
VBCV (S = 3)	24	70.0 (51.1, 85.9)	98.1	69.4 (47.0, 87.8)	95.7
VBCV (S = 5)	24	74.0 (55.5, 88.8)	99.4	65.8 (43.7, 84.7)	92.2
VBCV (S = 10)	24	78.0 (60.2, 91.6)	99.8	62.5 (40.9, 81.8)	87.5
BS (B = 100)	24	90.0 (75.6, 98.2)	>99.9	58.7 (38.5, 77.5)	80.3
BS (B = 1000)	24	94.0 (82.1, 99.5)	>99.9	52.1 (32.5, 71.3)	58.2
JK (S = 100)	24	86.0 (70.3, 96.4)	>99.9	65.9 (45.4, 83.7)	93.8
JK (S = 1000)	24	90.0 (75.9, 98.2)	>99.9	58.7 (38.5, 77.5)	80.3

CrI = credible interval.

provide the most reliable estimate of the true test performance of the tuned model?

To determine which resampling strategies most reliably approximate true test performance, we quantified the Bias [2] between validation and external test Macro F1 values for each method and compared them using a Bayesian Bradley-Terry framework (Table S16–S27; full model output in Table S28). To ensure that our evaluation reflected the full range of realistic model-selection behavior, we assessed the full range of possible LVs, using the 50% training set from each of the 10 splits to compute validation Macro F1 from each resampling method and the remaining data to compute testing Macro F1. By evaluating all LVs we captured the full spectrum of models practitioners may adopt in applied settings.

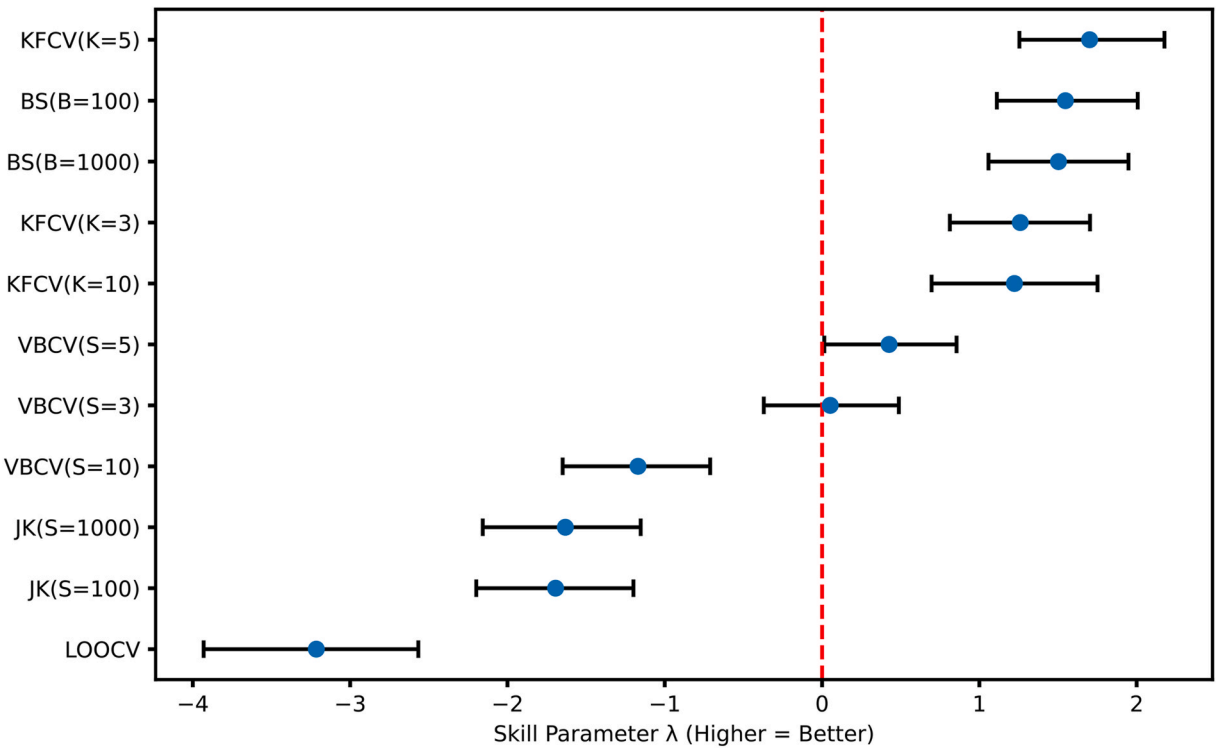


Fig. 3. The posterior mean skill parameters (λ) for each resampling strategy, along with their 95% credible intervals. In this framework, higher λ values indicate a greater probability that a method will produce lower validation-test bias relative to competing approaches. A value of $\lambda = 0$ represents an average-performing method, while positive or negative deviations indicate above- or below-average reliability, respectively.

Within the Bradley-Terry model, methods were compared through pairwise "wins," defined as instances in which a resampling strategy produced a validation Macro F1 closer to the true external test Macro F1 than a competing method under identical study conditions [53]. Because this model was fit in a fully Bayesian framework using MCMC, it generated posterior distributions over each method's latent "skill" in approximating test performance. This probabilistic ranking allows resampling strategies to be interpreted not merely by point estimates of bias, but by their overall reliability in producing validation metrics that generalize to independent data.

Following this approach, Fig. 3 presents the posterior mean skill parameters and corresponding 95% CrIs. KFCV ($K = 5$) exhibited the highest skill ($\lambda = 1.70$, 95% CrI: 1.28-2.16), indicating it most consistently minimized validation-test bias across comparisons. Bootstrap methods also performed strongly, with BS ($B = 100$) and BS ($B = 1000$) showing similar skill ($\lambda \approx 1.50$-1.55) and overlapping credible intervals, suggesting comparable reliability. Other K-fold approaches ($K = 3$ and $K = 10$) demonstrated moderate performance ($\lambda \approx 1.22$-1.26), whereas all VBCV, jackknife, and LOOCV methods showed substantially lower or negative skill values. In particular, LOOCV yielded the lowest skill

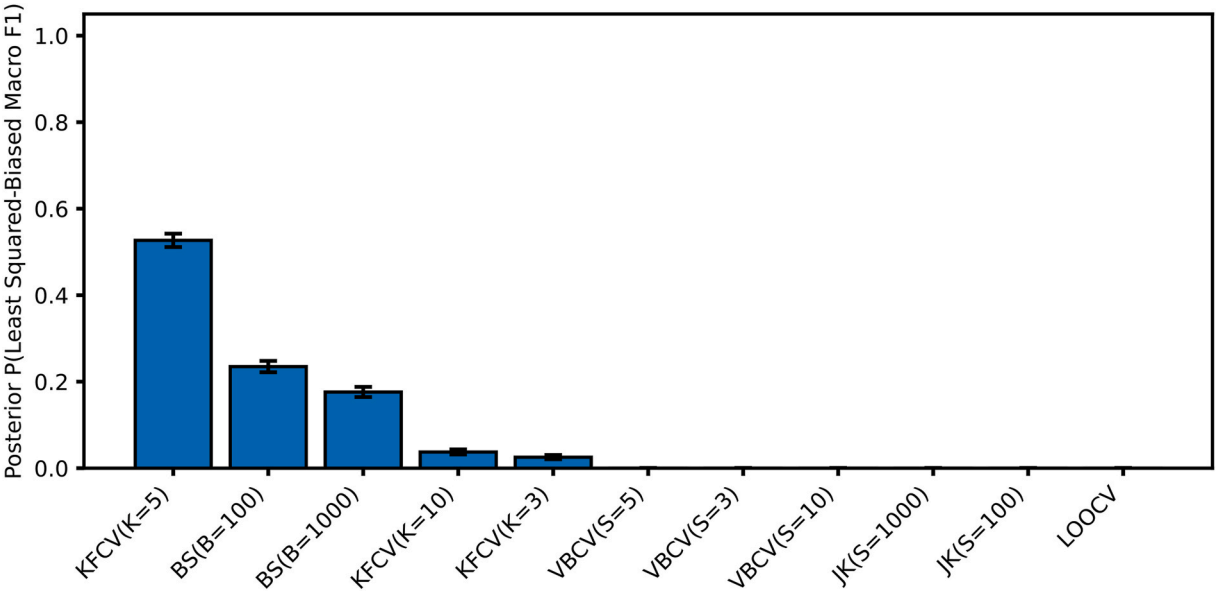


Fig. 4. Mean posterior probability and 95% CrIs that a method generates the least squared biased Macro F1 score among all methods.

($\lambda = -3.22$), indicating a consistently higher likelihood of producing biased estimates relative to other strategies.

Because pairwise performance is governed by $P(i > j) = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}}$ (equation (2)), for a given pair (i, j), these differences in λ translate directly into probabilistic advantages between methods. Therefore, one can estimate the average probability that KFCV (K = 5) will produce less squared bias than KFCV (K = 3) by plugging in equation (2) for i and j, respectively. Doing so should return a P (i > j) of ~62%. Which means KFCV (K = 5) is almost twice as likely to outcompete KFCV (K = 3) in terms of a least squares bias Macro F1 score during validation.

Furthermore, the poor performance of VBCV, LOOCV, and jackknife methods can be explained by the fact that all resampling was conducted at the sample level rather than the spectra level, which severely limits the effective validation set size in each split. In LOOCV, each validation fold contains only a single sample, meaning the validation Macro F1 is effectively constrained to either 0 or 1 depending on whether that sample is correctly classified, producing an extremely high-variance and discretized estimate of performance. This leads to systematic mismatch with the external test Macro F1, which is computed over many samples and is therefore much more stable. Similarly, jackknife approaches, particularly with larger deletion sizes, remove substantial portions of the data during training, resulting in models that are fit on reduced and potentially unrepresentative sample sets, which inflates bias when compared to full-data test performance. VBCV further exacerbates this issue by enforcing deterministic, ordered splits that may not preserve class balance at the sample level, leading to validation folds that are not representative of the underlying class distribution. Collectively, these factors cause validation estimates from LOOCV, jackknife, and VBCV to be either overly discrete, high-variance, or systematically unrepresentative, resulting in the large squared biases reflected in their strongly negative skill parameters.

Fig. 4 shows the posterior probability that each method is the single best estimator of test performance, defined as the proportion of posterior samples in which a method attains the highest skill. KFCV (K = 5) had the highest probability of being optimal (P ≈ 0.53, 95% CrI: 0.51–0.54), indicating that it was the best-performing method in over half of posterior draws. Bootstrap methods provided secondary support, with BS (B = 100) and BS (B = 1000) achieving probabilities of approximately 0.23 and 0.18, respectively, reflecting consistent but not dominant performance. All remaining methods exhibited negligible posterior probability ($\sim 10^{-4}$), indicating that they almost never outperformed the top methods across the posterior distribution. While the skill parameter reflects the relative strength of a method in pairwise comparisons, the posterior probability of being best reflects its ability to dominate all alternatives simultaneously; together, these metrics show that KFCV (K = 5) is both the strongest and most consistently optimal resampling strategy for estimating test Macro F1 in this dataset, whereas LOOCV, jackknife, and VBCV approaches are unlikely to provide reliable estimates of true test performance.

Notably, the observation that KFCV outperforms bootstrap methods is somewhat unexpected, as bootstrap resampling is often considered more suitable for estimating generalization error due to its ability to approximate sampling variability and utilize multiple resampled training sets [48]. One explanation is that all resampling was performed at the sample level, and bootstrap sampling with replacement can lead to repeated inclusion of the same samples and omission of others, effectively reducing the diversity and representativeness of individual training sets. This can introduce additional variance and bias in estimated performance, particularly for small to moderate sample sizes typical of spectroscopic datasets. In contrast, K-fold cross-validation ensures that each sample is used exactly once for validation and K−1 times for training, producing more balanced and representative splits that better reflect the full dataset distribution. As a result, KFCV (K = 5) may provide a more stable and less biased estimate of test Macro F1 under these conditions, despite the general expectation that bootstrap methods perform well for error estimation.

Overall, the Bayesian Bradley-Terry analysis demonstrates that resampling strategy has a substantial impact on the accuracy of test performance estimation for PLS-DA models. K-fold cross-validation, particularly KFCV (K = 5), consistently provided the most reliable and least biased estimates, outperforming bootstrap approaches despite their typical use for error estimation. In contrast, LOOCV, jackknife, and VBCV methods produced more biased estimates due to small or unrepresentative validation sets when applied at the sample level. These findings highlight that balanced, representative resampling schemes are critical for obtaining stable and generalizable performance estimates in spectroscopic machine learning applications.

4. Limitations

Although this study rigorously evaluates multiple resampling strategies across twelve diverse spectral datasets, several limitations should be acknowledged. First, the dataset sizes, although typical of many spectroscopic studies, are modest at the level of biological or specimen replicates, even when the number of spectra per sample is large. Machine learning literature consistently demonstrates that resampling-based error estimates become noisy and unstable when the number of independent samples is small, particularly for leave-one-out and high-fold CV schemes [46,49]. In chemometrics, the challenge is exacerbated because replicate spectra from the same biological specimen are not statistically independent, which can inflate apparent performance and distort variance estimates if not carefully controlled [46]. For example, the Holman and Kurouski (2023) study had the highest number of spectra per sample (fifty) and consequently performed the best during test accuracy estimation. While our study mitigated this risk by restricting model tuning and resampling to sample-level partitions, the fundamental constraints of small-sample spectral datasets still limit the generalizability of the reported conclusions. Future work with substantially larger, prospectively designed datasets would permit more reliable bias-variance decomposition and facilitate validation of resampling behavior under higher-power conditions.

Additionally, while the inclusion of multiple independent studies improves robustness, the total evidence base remains modest, and the conclusions should therefore be interpreted as indicative rather than definitive. Larger, prospectively designed datasets with greater numbers of independent biological or experimental samples will be necessary to confirm these findings and more fully characterize behavior under higher statistical power.

Finally, although Bayesian methods offer rich characterization of uncertainty and outperform frequentist testing for small-n scenarios, our Bayesian analyses rely on modeling assumptions that introduce their own limitations. The posterior probabilities, for example, summarize evidence across a limited set of nine studies that share common experimental and preprocessing philosophies. Accordingly, the Bayesian analyses are intended as exploratory, comparative summaries of resampling behavior rather than definitive or universal rankings. The Beta-Binomial framework preserves interpretability but simplifies model behavior into binary “non-harmful vs. harmful” outcomes, thereby discarding granularity in Macro F1 changes. Similarly, the Bradley-Terry model assumes transitivity and the existence of a latent skill parameter for each resampling method; assumptions that, while standard, may not fully represent complex interactions between dataset structure, spectral modality, and parameter-selection behavior [61]. Finally, although all chains achieved convergence diagnostics indicative of reliable sampling, Bayesian methods cannot compensate for intrinsic limitations in dataset diversity or size; the posterior probabilities reported here should therefore be interpreted as evidence conditional on our specific study corpus rather than as universal rankings. Broader validation across laboratories, instruments, and specimen types will be essential to establish externally generalizable guidelines for SML resampling practices.

5. Practical guidelines

Although the present study focuses on comparing resampling strategies for PLS-DA models, the results also provide several practical insights that may guide future SML study design. Across the twelve datasets examined, we observed that some models were essentially insensitive to the choice of resampling method (e.g., Tables S6, S11, S13, and S15), whereas others showed substantial variability in both validation curves and selected latent variables (e.g., Tables S8 and S12). This divergence is best understood by considering the relationship between within-sample replicate homogeneity and between-sample independent variation, a well-recognized principle in chemometric sampling theory [62].

When replicate spectra are highly consistent within each sample, resampling tends to remove only redundant information, yielding nearly identical model performance regardless of the validation scheme used. In contrast, when classes are represented by relatively few biological or specimen-level samples (or when those samples vary substantially) resampling partitions necessarily capture different aspects of the class structure, resulting in greater variability in test accuracy and LV selection. This behavior is not a flaw of resampling, but rather a reflection of the dataset's statistical structure.

Taken together, our findings support three practical recommendations for spectral machine learning practitioners:

1. ***Prioritize the number of independent samples per class over the number of technical replicates.*** This aligns with established chemometric principles that emphasize representativity and sample-level variance over dense replicate acquisition [62]. Technical replicates quickly reach diminishing returns, whereas additional biological or specimen-level samples enhance model generalizability. In general, 5-10 technical replicates is sufficient [62].
2. ***Use resampling stability as a diagnostic tool.*** When different resampling methods return similar accuracies and LV choices, the dataset likely contains sufficient independent sample variation [63]. When results diverge substantially across methods, this signals that the dataset may be underpowered at the sample level or that class structure is unstable.
3. ***Interpret LV sensitivity in the context of data structure.*** Large fluctuations in LV selection do not necessarily indicate algorithmic instability; they may instead reflect insufficient or unbalanced sampling, a phenomenon also noted in cross-validation literature [48, 63].

Additionally, several other common issues—although not directly addressed in this study—were identified during dataset acquisition, including:

1. **Misuse of preprocessing during validation.** Certain preprocessing techniques such as mean and median centering and multiplicative scatter correction rely on the available data to normalize each spectrum. If used, they should be computed within each training fold only and then applied to the corresponding validation/test data; performing these operations globally prior to resampling allows information from validation samples to influence model training leading to optimistically biased performance estimates [64]. Avoiding these steps altogether and switching to area or vector normalizations are another safe alternative.
2. **Diagnostics at the spectra-level rather than sample-level.** Because multiple spectra collected from the same specimen are highly correlated, treating them as independent observations inflates effective sample size and leads to overly optimistic accuracy, variance, and uncertainty estimates. Therefore, model evaluation should restrict train and validation/test data at the sample-level, ensuring no samples' spectra appear on both sides.
3. **Unclear validation-testing framework.** As one can see from Table 2, a majority of the studies did not discuss which resampling method was used during validation and some did not even discuss whether they used validation. Under these circumstances, it can become difficult for readers to trust the reliability of reported results when our data suggests dramatic differences across resampling methods for given purposes. Thus, the type of resampling method should be clearly stated.
4. **Data availability.** Finally, limited data availability continues to hinder reproducibility and method validation in SML studies. Many published works do not provide raw spectra, metadata, or preprocessing details, preventing independent verification and benchmarking of models. In Karl Popper's (translated) words, "single occurrences that cannot be reproduced are of no significance to science." [65] To address this, datasets should be deposited in accessible repositories with clear documentation of preprocessing, partitioning, and modeling procedures. For example, Zenodo and Dryad are two of many free public repositories to upload and share data and other information from studies. Transparent data sharing enables reproducibility, facilitates method comparison, and strengthens the evidentiary value of reported models [66].

By presenting these principles, we aim to provide users with practical heuristics for evaluating dataset adequacy and anticipating how resampling strategies might interact with their data. Ultimately, increasing the number of independent samples, coupled with sufficient numbers of technical replicates, reduces the likelihood that study design will inadvertently influence model performance.

6. Conclusions

This tutorial study demonstrates that resampling strategy plays a critical and often underappreciated role in both model selection and performance estimation in spectroscopic machine learning. While bootstrap and jackknife methods provide strong stability in preserving predictive performance, they do so at the expense of model parsimony, consistently favoring more complex models. In contrast, LOOCV and selected Venetian blinds and 3 and 10-fold cross validation configurations offer a more balanced compromise, maintaining performance while encouraging simpler model structures. Importantly, these results highlight that commonly grouped approaches such as K-fold cross-validation are not interchangeable; specific parameter choices materially influence reliability.

When evaluating the ability of resampling methods to estimate true test performance, K-fold cross-validation (particularly with 5 folds) consistently provided the most accurate and least biased estimates. In contrast, LOOCV, jackknife, and Venetian blinds approaches produced higher bias due to small or unrepresentative validation sets when applied at the sample level. Taken together, these findings emphasize that resampling strategy should be treated as a tunable component of the modeling pipeline rather than a fixed default. Practitioners should prioritize methods that balance representativeness and stability, as these characteristics are essential for obtaining reliable, generalizable performance estimates in applied spectroscopic machine learning.

CRediT authorship contribution statement

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

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Dmitry Kurouski reports financial support was provided by Texas A&M University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2026.345655.

Data availability

The Python scripts used throughout data analysis can be found here: https://github.com/crownboxco/Python-Scripts-for-Cross-Study-Validation-of-Resampling-Methods. All relevant datasets are publicly available and cited in Table 1.

References

[1] M.-A. Belabbas, P.J. Wolfe, Spectral methods in machine learning and new strategies for very large datasets, Proc. Natl. Acad. Sci. 106 (2) (2009) 369–374.
[2] J. Zeng, Y. Guo, Y. Han, Z. Li, Z. Yang, Q. Chai, W. Wang, Y. Zhang, C. Fu, A review of the discriminant analysis methods for food quality based on near-infrared spectroscopy and pattern recognition, Molecules 26 (3) (2021) 749.
[3] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Near-infrared (NIR) spectroscopy for motor oil classification: from discriminant analysis to support vector machines, Microchem. J. 98 (1) (2011) 121–128.
[4] D. de Carvalho Lopes, A.J.S. Neto, Classification and authentication of plants by chemometric analysis of spectral data, in: Comprehensive Analytical Chemistry, vol. 80, Elsevier, 2018, pp. 105–125.
[5] N. Blake, R. Gaifulina, L.D. Griffin, I.M. Bell, G.M. Thomas, Machine learning of Raman spectroscopy data for classifying cancers: a review of the recent literature, Diagnostics 12 (6) (2022) 1491.
[6] M.T. Brown, L.R. Wicker, Discriminant analysis, in: Handbook of Applied Multivariate Statistics and Mathematical Modeling, Elsevier, 2000, pp. 209–235.
[7] A.M. Molinaro, R. Simon, R.M. Pfeiffer, Prediction error estimation: a comparison of resampling methods, Bioinformatics 21 (15) (2005) 3301–3307.
[8] H. Nawaz, F. Bonnier, P. Knief, O. Howe, F.M. Lyng, A.D. Meade, H.J. Byrne, Evaluation of the potential of Raman microspectroscopy for prediction of chemotherapeutic response to cisplatin in lung adenocarcinoma, Analyst 135 (12) (2010) 3070–3076.
[9] T. Jaki, T.-L. Su, M. Kim, M.L. Van Horn, An evaluation of the bootstrap for model validation in mixture models, Commun. Stat. Simulat. Comput. 47 (4) (2018) 1028–1038.
[10] A.A. Christy, S. Kasemsumran, Y. Du, Y. Ozaki, The detection and quantification of adulteration in olive oil by near-infrared spectroscopy and chemometrics, Anal. Sci. 20 (6) (2004) 935–940.
[11] A. Borin, M.F. Ferrao, C. Mello, D.A. Maretto, R.J. Poppi, Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk, Anal. Chim. Acta 579 (1) (2006) 25–32.
[12] C. Lu, B. Xiang, G. Hao, J. Xu, Z. Wang, C. Chen, Rapid detection of melamine in milk powder by near infrared spectroscopy, J. Near Infrared Spectrosc. 17 (2) (2009) 59–67.
[13] G. Chen, X. Lin, D. Lin, X. Ge, S. Feng, J. Pan, J. Lin, Z. Huang, X. Huang, R. Chen, Identification of different tumor states in nasopharyngeal cancer using surface-enhanced Raman spectroscopy combined with Lasso-PLS-DA algorithm, RSC Adv. 6 (10) (2016) 7760–7764.
[14] R.P. Aguiar, E.T. Falcao, C.A. Pasqualucci, L. Silveira Jr., Use of Raman spectroscopy to evaluate the biochemical composition of normal and tumoral human brain tissues for diagnosis, Laser Med. Sci. 37 (1) (2022) 121–133.
[15] G.R. Lloyd, L.E. Orr, J. Christie-Brown, K. McCarthy, S. Rose, M. Thomas, N. Stone, Discrimination between benign, primary and secondary malignancies in lymph nodes from the head and neck utilising Raman spectroscopy and multivariate analysis, Analyst 138 (14) (2013) 3900–3908.
[16] G. Moreno-Martin, M.E. León-González, Y. Madrid, Simultaneous determination of the size and concentration of AgNPs in water samples by UV–vis spectrophotometry and chemometrics tools, Talanta 188 (2018) 393–403.
[17] L. Huang, X. Zhang, Z. Zhang, Sensor array for qualitative and quantitative analysis of metal ions and metal oxyanion based on colorimetric and chemometric methods, Anal. Chim. Acta 1044 (2018) 119–130.
[18] E.H. de Paulo, G.B. Magalhães, M.P. Moreira, M.H. Nascimento, O.A. Heringer, P. R. Filgueiras, M.F. Ferrão, Classification of water by bacterial presence using chemometrics associated with excitation-emission matrix fluorescence spectroscopy, Microchem. J. 197 (2024) 109804.
[19] M. Peterson, D. Kurouski, Non-destructive identification of dyes on fabric using near-infrared Raman spectroscopy, Molecules 28 (23) (2023) 7864.
[20] A.P. Holman, M. Peterson, E. Linhart, D. Kurouski, Using surface-enhanced Raman spectroscopy to probe artificial dye degradation on hair buried in multiple soils for up to eight weeks, Sci. Rep. 14 (1) (2024) 6469.
[21] A.P. Holman, D. Kurouski, Surface-enhanced Raman spectroscopy enables confirmatory detection of dyes on hair submerged in hypolimnion water for up to twelve weeks, J. Forensic Sci. (2023).
[22] A.P. Holman, D. Kurouski, Role of Race/Ethnicity, sex, and age in surface-enhanced raman spectroscopy-and infrared spectroscopy-based analysis of artificial colorants on hair, ACS Omega (2023).
[23] A.P. Holman, A. Maalouf, D. Kurouski, DyeSPY: establishing the first forensic SERS reference for hair dye colorant evidence, Anal. Chem. (2025).
[24] N.K. Goff, J.F. Guenther, J.K. Roberts III, M. Adler, M.D. Molle, G. Mathews, D. Kurouski, Non-invasive and confirmatory differentiation of hermaphrodite from both male and female cannabis plants using a hand-held Raman spectrometer, Molecules 27 (15) (2022) 4978.
[25] N.K. Goff, J.F. Guenther, J.K. Roberts III, M. Adler, M.D. Molle, G. Mathews, D. Kurouski, Dataset for: non-invasive and confirmatory differentiation of hermaphrodite from both Male and female cannabis plants using a. Hand-Held Raman Spectrometer, Zenodo, 2026.
[26] S. Higgins, V. Serada, B. Herron, K.R. Gadhave, D. Kurouski, Confirmatory detection and identification of biotic and abiotic stresses in wheat using Raman spectroscopy, Front. Plant Sci. 13 (2022) 1035522.
[27] S. Higgins, V. Serada, B. Herron, K.R. Gadhave, D. Kurouski, Dataset For: Confirmatory Detection and Identification of Biotic and Abiotic Stresses in Wheat Using Raman Spectroscopy, Zenodo, 2026.
[28] A.P. Holman, D. Kurouski, Dataset For: Role of Race/Ethnicity, Sex, and Age in Surface-Enhanced Raman Spectroscopy- and Infrared Spectroscopy-based Analysis of Artificial Colorants on Hair, Zenodo, 2026.
[29] A.P. Holman, D.N. Pickett, A.E. Orr, A.M. Tarone, D. Kurouski, A nondestructive technique for the sex identification of third instar Cochliomyia macellaria larvae, J. Forensic Sci. 69 (6) (2024) 2075–2081.
[30] A.P. Holman, D.N. Pickett, A.E. Orr, A.M. Tarone, D. Kurouski, Dataset For: a Nondestructive Technique for the Sex Identification of Third Instar Cochliomyia macellaria Larvae, Zenodo, 2026.
[31] A. Rodriguez, B.B. Barcenilla, E. Hall, I. Kundel, A. Meyers, S. Wyatt, D. Shippen, D. Kurouski, Raman spectroscopy as a tool for assessing plant growth in space and on lunar regolith simulants, npj Microgravity 11 (1) (2025) 19.
[32] A. Rodriguez, B.B. Baarcenilla, E. Hall, I. Kundel, A. Meyers, S. Wyatt, D. Shippen, D. Kurouski, Dataset For: Raman Spectroscopy as a Tool for Assessing Plant Growth in Space and on Lunar Regolith Simulants, Zenodo, 2026.
[33] A.P. Holman, D.N. Pickett, H. West, A.M. Tarone, D. Kurouski, Portable fourier-transform infrared spectroscopy and machine learning for sex determination in third instar Chrysomya rufifacies larvae, J. Forensic Sci. (2025).
[34] A.P. Holman, D.N. Pickett, H. West, A.M. Tarone, D. Kurouski, Dataset For: Portable Fourier-transform Infrared Spectroscopy and Machine Learning for Sex Determination in Third Instar Chrysomya rufifacies Larvae, Zenodo, 2026.
[35] I.D. Juárez, A.P. Holman, E.J. Horn, A.S. Rogovskyy, D. Kurouski, External validation of raman spectroscopy for Lyme disease diagnostics, J. Biophot. 18 (5) (2025) e202400520.
[36] I.D. Juárez, A.P. Holman, E.J. Horn, A.S. Rogovskyy, D. Kurouski, Dataset For: External Validation of Raman Spectroscopy for Lyme Disease Diagnostics, Zenodo, 2026.
[37] C. Sasaki, S. Bober, A.P. Holman, D. Kurouski, Identification of dyes on fabric exposed to lake and ocean water using near-infrared excitation Raman spectroscopy, Anal. Methods (2026).
[38] C. Sasaki, S. Bober, A.P. Holman, D. Kurouski, Dataset For: Identification of Dyes on Fabric Exposed to Lake and Ocean Water Using near-infrared Excitation Raman Spectroscopy, Zenodo, 2026.
[39] F. Kosmowski, T. Worku, Evaluation of a miniaturized NIR spectrometer for cultivar identification: the case of barley, chickpea and sorghum in Ethiopia, PLoS One 13 (3) (2018) e0193620.
[40] F. Muthreich, B. Zimmermann, H.J.B. Birks, C.M. Vila-Viçosa, A.W. Seddon, Chemical variations in Quercus pollen as a tool for taxonomic identification: implications for long-term ecological and biogeographical research, J. Biogeogr. 47 (6) (2020) 1298–1309.
[41] F.Z. Muthreich, Boris, Carlos M. Vila-Viçosa, H.John B. Birks, Alistair W.R. Seddon, Chemical Variations in Quercus Pollen as a Tool for Taxonomic Identification: Implications for long-term Ecological and Biogeographical Research, 2020.
[42] A. Banerjee, A. Gokhale, R. Bankar, V. Palanivel, A. Salkar, H. Robinson, J. S. Shastri, S. Agrawal, G. Hartel, M.M. Hill, Rapid classification of COVID-19 severity by ATR-FTIR spectroscopy of plasma samples, Anal. Chem. 93 (30) (2021) 10391–10396.
[43] A.G. Arghya Banerjee, Renuka Bankar, Viswanthram Palanivel, Akanksha Salkar, Harley Robinson, Jayanthi S. Shastri, Sachee Agrawal, Gunter Hartel, Michelle M. Hill, Sanjeeva Srivastava, Rapid Classification of COVID-19 Severity by ATR-FTIR Spectros-copy of Plasma Samples, 2021.
[44] A. Barney, V. Trojan, R. Hrib, A. Newland, J. Halámek, L. Halámková, From spectra to signatures: detecting fentanyl in human nails with ATR–FTIR and machine learning, Sensors 25 (1) (2025) 227.
[45] A. Barney, V. Trojan, R. Hrib, A. Newland, J. Halámek, L. Halámková, Dataset For: from Spectra to Signatures: Detecting Fentanyl in Human Nails with ATR–FTIR and Machine Learning, 2026.
[46] J.A. Westerhuis, H.C. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J. van Velzen, J. P. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (1) (2008) 81–89.
[47] U.G. Indahl, H. Martens, T. Næs, From dummy regression to prior probabilities in PLS-DA, J. Chemometr.: A J. Chemometrics Soc. 21 (12) (2007) 529–536.
[48] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer series in statistics, New-York, 2009.

[49] S. Arlot, A. Celisse, A Survey of cross-validation Procedures for Model Selection, 2010.
[50] A.L. Pomerantsev, O.Y. Rodionova, A new method for soft probabilistic discrimination of ranked classes, Microchem. J. 212 (2025) 113329.
[51] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Ijcai, vol. 14, 1995, pp. 1137–1145. Montreal, Canada.
[52] B. Efron, R. Tibshirani, Improvements on cross-validation: the 632+ bootstrap method, J. Am. Stat. Assoc. 92 (438) (1997) 548–560.
[53] J. Wainer, A Bayesian Bradley-Terry model to compare multiple ML algorithms on multiple data sets, J. Mach. Learn. Res. 24 (341) (2023) 1–34.
[54] N. Japkowicz, M. Shah, Performance evaluation in machine learning, in: Machine Learning in Radiation Oncology: Theory and Applications, Springer, 2015, pp. 41–56.
[55] I. Alarab, S. Prakoonwit, Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques, Data Sci. Manage. 5 (2) (2022) 66–76.
[56] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1) (2002) 67–82.
[57] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, 2013.
[58] C. Beleites, R. Baumgartner, C. Bowman, R. Somorjai, G. Steiner, R. Salzer, M. G. Sowa, Variance reduction in estimating classification error using sparse datasets, Chemometr. Intell. Lab. Syst. 79 (1–2) (2005) 91–100.
[59] A.R. Flanagan, F.G. Glavin, A systematic review of multi-class and one-vs-rest classification techniques for near-infrared spectra of crop cultivars, in: 2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS), IEEE, 2023, pp. 1–8.
[60] U. Wickenberg-Bolin, H. Göransson, M. Fryknäs, M.G. Gustafsson, A. Isaksson, Improved variance estimation of classification performance via reduction of bias caused by small sample size, BMC Bioinf. 7 (1) (2006) 127.
[61] F. Caron, A. Doucet, Efficient Bayesian inference for generalized Bradley–Terry models, J. Comput. Graph Stat. 21 (1) (2012) 174–196.
[62] K.H. Esbensen, P. Geladi, Principles of proper validation: use and abuse of re-sampling for validation, J. Chemometr. 24 (3-4) (2010) 168–187.
[63] G. Varoquaux, Cross-validation failure: small sample sizes lead to large error bars, Neuroimage 180 (2018) 68–77.
[64] M.A. Bouke, S.A. Zaid, A. Abdullah, Implications of Data Leakage in Machine Learning Preprocessing: a multi-domain Investigation, 2024.
[65] U. Dirnagl, Rethinking research reproducibility, EMBO J. 38 (2) (2019). EMBJ2018101117.
[66] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (1) (2016) 1–9.

Dr. Dmitry Kurouski received his M.Sc. in Biochemistry from Belarussian State University in 2007. After earning his Ph.D. in Analytical Chemistry from The State University of New York at Albany in 2013, Dr. Kurouski joined the Chemistry Department at Northwestern University where he worked as a postdoctoral fellow in the laboratory of Richard P. Van Duyne. Prior to Texas A&M University, Dr. Kurouski worked as a Senior Research Scientist at BoehringerIngelheim Pharmaceuticals. His research interests are focused on disease diagnostics in a large spectrum of biological species and elucidation of the underlying molecular causes of neurodegenerative diseases. His group pioneers the nanoscale analysis of photocatalytic processes using tip-enhanced Raman spectroscopy and atomic force microscopy Infrared spectroscopy.

Aidan Holman is a graduate researcher in Dr. Dmitry Kurouski's group at Texas A&M AgriLife Research. Since beginning his PhD program in Toxicology in 2024, Aidan has remained focused on developing diagnostic tools for insects, plants, medicine, and forensics, that advantage spectroscopy and machine learning. His current interests include developing surface-enhanced Raman substrates for drug and trace evidence discovery to help assist law enforcement.