Analyst

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6an00346j

The infinite-dimensional nature of spectroscopy and why models succeed, fail, and mislead

Umberto Michelucci *[a,c] and Francesca Venturini [b,c]

Machine learning (ML) models have achieved strikingly high accuracies in spectroscopic classification tasks, often without a clear proof that those models used chemically meaningful features. Existing studies have linked these results to data preprocessing choices, noise sensitivity, and model complexity, but no unifying explanation is available so far. In this work, we show that these phenomena arise naturally from the intrinsic high dimensionality of spectral data. Using a theoretical analysis grounded in the Feldman–Hájek theorem and the concentration of measure, we show that even infinitesimal distributional differences, caused by noise, normalisation, or instrumental artefacts, may become perfectly separable in high-dimensional spaces. Through a series of specific experiments on synthetic and real fluorescence spectra, we illustrate how models can achieve near-perfect accuracy even when chemical distinctions are absent, and why feature-importance maps may highlight spectrally irrelevant regions. We provide a rigorous theoretical framework, confirm the effect experimentally, and conclude with practical recommendations for building and interpreting ML models in spectroscopy.

Received 29th March 2026,
Accepted 6th April 2026

DOI: 10.1039/d6an00346j

rsc.li/analyst

1 Introduction

Spectroscopy is the study of how matter interacts with electromagnetic radiation, typically by measuring the intensity of emitted light as a function of wavelength or frequency. The analysis of these spectral signatures non-destructively can reveal the composition and then chemical structure of a sample.

Extracting chemical and physical information from spectra is usually complex and requires involuted data processing pipelines that include steps such as, for example, baseline subtraction or smoothing.[1–3] Interpreting the output of machine learning (ML) models applied to spectra presents several challenges: the high number of wavelengths complicates interpretation, complex models might capture nonlinear interactions, making it difficult to connect features in spectra with chemical information about the sample (models are often black-boxes, for lack of interpretability). Complex models may fit noise rather than signal, making predictions useless.[4] Furthermore, it has become clear to the research community that attributing a prediction to specific wavelength bands does not have a unique solution, and different methods lead to different attributions.[5] This is due to the fact that explainability approaches measure how a specific model responds to changes in intensity at individual wavelengths, which very often include regions far from chemically significant peaks that the model learns to exploit due to subtle statistical differences.[6] Zehtabvar *et al.*[7] have found that data normalisation has a strong influence on the accuracy of ML models, something that seems strange, since data normalisation does not have a relationship with physico-chemical information about measurements. Contreras *et al.*,[6] clearly show how feature importance algorithms are susceptible to noise-induced fluctuations, although they fail to give a good explanation for this observation. They also note that because of the high dimensionality of the data, interpretation of the results is challenging.

Steinmann *et al.*[8] studied the problem of spurious correlation in a long article and compared it to the "clever Hans behaviour". The term *Clever Hans* comes from animal psychology, named after the horse Hans that apparently had learnt to understand human language. Hans (the horse) instead learnt to rely on the facial expressions of humans asking questions and was unable to give correct answers when not seeing the human face (for more interesting details on the case, you can read[9]). To paraphrase Steinmann *et al.*, in some cases, ML algorithms in spectroscopy are as dumb as a horse.† The fact that ML is seemingly capable of classifying any spectra dataset with a high accuracy has sparked the appearance of uncountable articles that use ML to extract the elusive chemical or

[a]*Lucerne University of Applied Sciences and Arts, Computer Science Department, 6343 Risch-Rotkreuz, Switzerland. E-mail: umberto.michelucci@hslu.ch*
[b]*Institute of Applied Mathematics and Physics, ZHAW Zurich University of Applied Sciences, Winterthur, 8400, Switzerland. E-mail: vent@zhaw.ch*
[c]*TOELT LLC, Research and Development, Duebendorf, Switzerland*

† The authors do not want to imply that horses are not intelligent animals, only that they cannot classify spectra accurately.

physical parameters sought without the need of involuted data analysis pipelines (see for an overview[10]).

How such models really generalise to new measurement setups or datasets is an open question that cannot be answered uniquely today. This depends on whether a model learns from physically meaning features (*e.g.*, an absorption or emission line), or from artefacts of the measurement process (*e.g.*, from the noise introduced by a specific electronics[11]). In the first case, the model is likely to generalise to new measurements well, whereas in the latter case the model simply overfits specific characteristics of the measurement apparatus, rendering it unreliable.

This article, for the first time, explains why the high dimensionality of spectroscopy (the number of intensity values in spectra is usually of the order of 10^3) is responsible for the ability of ML to classify almost all kinds of spectroscopy dataset, even in situations where the data itself contain no discernible feature to distinguish between classes (*e.g.*, all intensities in spectra in one class noticeably higher than in another class). Fundamentally, this work demonstrates that the effectiveness of ML models when applied to spectroscopy may be due in many cases to the high dimensionality of the spectral data rather than chemical–physical spectral features. Especially flexible models, such as random forests, can obtain an almost perfect classification accuracy even using spectral regions that do not contain relevant physico-chemical features because of the high dimensionality of the data. This work demonstrates that, in high-dimensional spaces, even subtle differences in the statistical properties of the signal across classes can enable a model to achieve seemingly perfect classification accuracy, despite the spectra themselves lacking sufficient information to justify such performance. This may be particularly relevant in reflectance and fluorescence spectroscopy, where the spectra usually have broad features rather than sharp specific signatures, such as in Raman-spectroscopy.

The contributions of this article are the following. (i) We present a mathematical discussion of the role of high dimensionality in finite- and infinite-dimensional cases, with a specific discussion of how to translate those abstract results to spectroscopy. (ii) We present a series of experiments on synthetic data to show under which conditions high-dimensionality becomes relevant in classification. (iii) We show how this phenomenon appears in real fluorescence spectroscopy data. (iv) Finally, we describe how spectroscopists should change their way of analysing spectra to take into account the effect of high dimensionality.

This article is structured in the following way. In section 2 we discuss the mathematical reasons for the behaviour of Gaussian distributed data in finite- and infinite-dimensional spaces (the Feldman–Hájek theorem[12,13]). We then proceed to generalise the findings for the non-Gaussian distributed data and explain why it is applicable to spectroscopy data. In section 3 we describe a series of experiments on synthetic data and real data used to investigate the effect of dimensionality. In section 4 we present the results obtained on synthetic and real data. In section 5, we discuss the relevance of the results for spectroscopy. Finally, in section 6 we discuss conclusions and limitations.

2 Methods

To study how the classification accuracy of models varies when applied to spectral data as a function of the number of intensities (referred to in this paper as the "spectrum" dimensions), it is useful to begin with the most basic case: data that follow a normal distribution. Later in this paper, we will expand the discussion to more realistic cases that reflect spectroscopy data more closely.

2.1 Gaussian case

The key mathematical theorem that describes under which conditions the data distributed according to two different Gaussians are separable (in other words, perfectly classifiable) is the Feldman–Hájek thorem, proved in 1958 by Feldman[12] and Hájek independently.[13] This theorem is well known in Gaussian measure theory, but less in applied machine learning circles. The theorem provides the conditions under which two Gaussian measures are either *mutually absolutely continuous* or *mutually singular* (to better understand the meaning of these two concepts, the reader is referred to ref. 14). In intuitive terms, it tells when two sets of data, X_1 and X_2, drawn from *different* multivariate normal distributions, can be regarded as perfectly separable from a measure-theory point of view (which means that there exists a possibly very complex classifier that can separate them with zero classification error).

The implications of the Feldman–Hájek theorem are highly relevant for spectroscopy. A spectrum can be viewed as a point in a high-dimensional space, where intensity at each wavelength (or pixel) represents one coordinate. When this dimensionality is large (for example, 10^3 intensities values), the *geometry* of the space in which spectra are defined changes dramatically. The theorem states that, under assumptions verified in the case of spectroscopy, in finite dimensions, two Gaussian distributions with slightly different means or variances always overlap to some extent and can never be perfectly classified. In contrast, in infinite (or with a good approximation in very high) dimensions, even the smallest difference in mean or covariance makes the two distributions *mutually singular*, meaning that they occupy disjoint regions of the space, and as such, they can be perfectly classified by an appropriate algorithm. For a spectroscopist, this provides a rigorous explanation for a common observation: ML models often achieve a very high accuracy even when spectra appear indistinguishable. The Feldman–Hájek theorem shows that this behaviour is a *geometric* consequence of high dimensionality: tiny instrumental artefacts, baseline shifts, or preprocessing differences can make two classes of spectra perfectly separable, even in the absence of any genuine physico-chemical distinction. Thus, the theorem clarifies why models may

"succeed" mathematically while not learning from physico-chemical meaningful information.

2.2 Non Gaussian case

Spectral data typically do not follow a normal (Gaussian) distribution. Consequently, one might initially assume that the previous discussion does not adequately describe the behaviour of real-world datasets. The Feldman–Hájek theorem can be generalized[15] to the case of an infinite countable collection of Gaussian mixture models (in what is called the Gaussian mixture dichotomy theorem). Thus, since the distribution of any data can be approximated to an arbitrarily high degree of accuracy by such mixtures, this extension effectively applies to essentially any dataset. That substantially means that in the limit of infinite dimensions we can expect an almost perfect classification for basically every dataset (up to certain limits that depend on the chosen classifier and dataset).

2.3 Concentration of measure

A further justification of the result comes from the phenomenon known as *concentration of measure*. To intuitively understand it, consider the following example. In our everyday 3-dimensional world, a solid orange is mostly "fruit" and very little "peel". If you were to thin the peel slightly, the amount of fruit inside would not change much. However, in high-dimensional spaces (such as the 1024-pixel space of a spectrometer), geometry works backward. An orange in 1000 dimensions (or more) will be almost entirely concentrated on the peel, and it will be almost completely empty inside (that would make peeling an orange quite an endevour, so we should probably be thankful not to live in 1000 dimensions).

Basically, as you add dimensions, the "volume" of a shape migrates away from the centre and traps itself almost entirely in the outer shell. In 1024 dimensions, a "solid" ball (your orange) is essentially empty; 99.9% of its contents exist only in a paper-thin layer on the surface. This means that if you pick a random point that is part of the ball, its distance from the centre will almost always be the same, since it will be with almost certainty on the shell (the orange peel) (this is something that defies completely intuition)! In other words, nearly all points of the ball have a norm $\|x\|_2$ (the length of the vector from the origin to a point x part of the ball) close to a typical value.

This phenomenon shows itself in high-dimensional spaces, and the probability mass (intuitively the values of the norm of the arrays $\|x\|_2 = ({x_1}^2 + \cdots + {x_n}^2)^{1/2}$) for a Gaussian distributed dataset in n dimensions tends to concentrate around $\sigma\sqrt{n}$ (assuming for simplicity that $x \sim \mathcal{N}(0, \sigma^2 I_n)$). For a visualisation of this phenomenon, in Fig. 1 we show the distributions of $\|x\|_2$ (the length of vectors x) sampled from two Gaussian distributions $\mathcal{N}(0, 1.0^2 I_n)$ and $\mathcal{N}(0, 1.1^2 I_n)$ where I_n is the identity matrix of dimensions $n \times n$. For illustrative reasons, we consider the case of isotropic covariances, but this phenomenon is still happening for generic covariances. In panel (A) of Fig. 1 it can be seen that in dimension 2, the $\|x\|_2$ distributions have a high overlap (as intuitively clear since the two

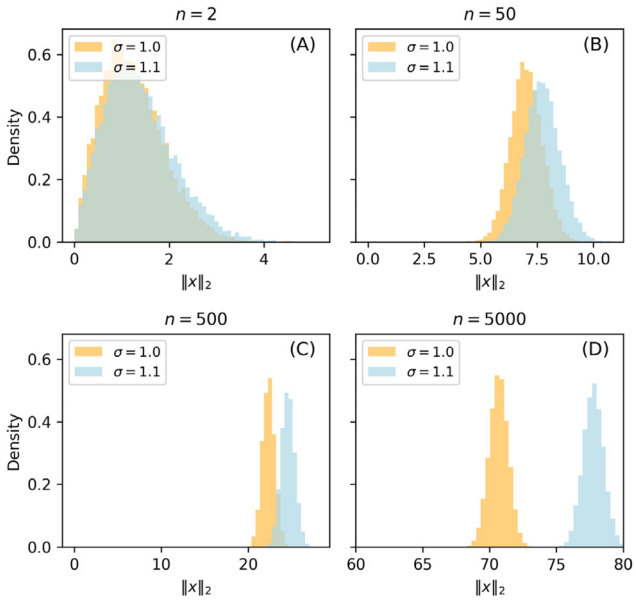

Fig. 1 Illustration of the concentration of measure for multivariate Gaussian distributions. Shown are the empirical distributions of $\|x\|_2$ for samples drawn from $\mathcal{N}(0, 1.0^2 I_n)$ (light blue) and $\mathcal{N}(0, 1.1^2 I_n)$ (yellow), for increasing dimensionalities n = 2, 50, 500, and 5000 (panels A–D). In low dimensions the two distributions overlap substantially, but as n increases the probability mass concentrates sharply around the typical radius $\sigma\sqrt{n}$, and even small variance differences cause almost complete separation. This illustrates how, in high-dimensional spaces, measures supported by Gaussian (and many non-Gaussian) distributions become effectively disjoint, providing an intuitive geometric basis for the Feldman–Hájek theorem and for the sensitivity of high-dimensional classifiers to minute statistical differences.

distributions have the same mean and only slightly different covariances). As the dimension increases (panel (B), (C), and (D)), distributions overlap decrease, until the dimensionality is high enough (panel (D)), and the two have almost no overlap anymore.

In general, random variables with finite variance exhibit extremely small relative fluctuations even when not Gaussian: most realisations (the measured values) lie very close to their expected value. This implies that the geometry of high-dimensional data is effectively governed by its first- and second-order statistics (mean and covariance) and that differences in these quantities dominate the behaviour of distances and overlaps between distributions. Hence, the Gaussian assumption underlying the Feldman–Hájek theorem remains a valid guide for understanding separability and equivalence of high-dimensional or averaged data, even when the underlying distributions deviate from strict normality.

2.4 Effect in Ohter fields

The "dimensional traps" identified in this work are not unique to spectroscopy; they represent a fundamental challenge across high-dimensional measurement sciences. In genomics, for instance, the $p \gg n$ problem (where the number of genes far exceeds the number of patients) frequently leads to 'phantom'

biomarkers[16,17] that fail to replicate in clinical trials—a phenomenon often attributed to the model's ability to find a separating hyperplane in the high-dimensional noise floor of the microarray or sequencing data. Similarly, in functional MRI (fMRI) neuroimaging, researchers have documented 'voodoo correlations'[18] (to use the same words that Vul *et al.* used) where high-dimensional voxel-wise patterns yield near-perfect classification of psychological states, only to be later identified as artefacts of head motion or instrumental sampling bias.

3 Experiments

To show convincingly that high-dimensionality may be responsible for the exceptional performance of ML models in spectroscopy, a series of experiments on synthetic datasets and on a real fluorescence dataset are presented in this article.

The overview of the experiments is presented in Table 1. Experiments are indicated with N1, N2, N3, and N4 for noise classification experiments, with S1, S2, and S3 for synthetic spectra classification experiments, and with R1a to R5b for real data classification experiments.

3.1 Gaussian noise classification

As a first set of experiments (N1, N2, and N3), we generated two classes of random noise arrays of dimension n, sampled from an isotropic multivariate distribution in the range $\mathcal{N}(\mu_1, {\sigma_1}^2 I_n)$ for class 1 and $\mathcal{N}(\mu_2, {\sigma_2}^2 I_n)$ for class 2. These experiments aim to testing the accuracy of classifiers in distinguishing the two classes for various values of the standard deviation σ_1 and σ_2 varying dimensionality n, with different approaches: QDA (experiments N1 and N3) and LDA decision boundary (N2). In N1 and N2 we have $\mu_1 = \mu_2 = 1$, while in N3 we have $\mu_1 = \mu_0 = 0$. Choosing the covariance matrix as $\sigma^2 \mathrm{I}_n$ is

Table 1 Overview of experiments and their core ideas

ID	Experiment	Experiment Details
N1	Gaussian noise: $\Delta\sigma$ sweep (QDA)	Classify white noise from two multivariate Gaussians with equal means ($\mu_1 = \mu_2 = 1$) and different isotropic and Toeplitz (with parameter $\rho = 0.95$) covariances; we varied the variance gap $\Delta\sigma$: = $\lvert\sigma_2 - \sigma_1\rvert$ from 0 to 2 while keeping $\sigma_1 = 1$ fixed. We have tested $n = \{5, 10, 50, 500\}$. We have measured classification accuracy
N2	Gaussian noise: Bayes "oracle" boundary	Classify white noise from two multivariate Gaussians with equal means ($\mu_1 = \mu_2 = 1$) and different isotropic covariances; we varied the variance gap $\Delta\sigma$: = $\lvert\sigma_2 - \sigma_1\rvert$ from 0 to 2 while keeping $\sigma_1 = 1$ fixed. We have tested $n = \{30, 100, 500, 1000, 5000\}$. We classify white noise by using the ideal threshold on $\sum(x_j - \mu)^2$
N3	Gaussian noise: accuracy *vs.* dimension n	Classify white noise from two multivariate Gaussians with equal means ($\mu_1 = \mu_2 = 0$) and different isotropic covariances; we varied the variance gap in $\Delta\sigma$: = $\lvert\sigma_2 - \sigma_1\rvert = \{0.1, 0.3, 0.6, 0.9, 1.2, 1.5, 2.0\}$ while keeping $\sigma_1 = 1$ fixed. We tested $n = \{1, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}$. We calculate the accuracy of QDA for various values of n.
N4	Skew-normal noise: 2D parameter sweeps	Classify two classes of white noise from multivariate skew-normal.19 For this experiment, we used the parameters $n = 50$, $\mu_1 = \mu_2 = 10$, $\sigma_1 = 1$, $\gamma_1 = 0.5$. We then varied $\Delta\sigma/\sigma$ from 0 to 2, $\Delta\mu/\mu$ from 0 to 0.15, and $\Delta\gamma_1/\gamma_1$ from 0 to 8
S1	Synthetic spectra: truly identical classes	Classify two classes of spectra each composed of one Lorentzian (both centres chosen according to a normal distribution $\mathcal{N}(50, 10^2)$) and an FWHM $\xi = 7$. We generated $N = 500$ spectra for each class. We choose $n = \{5, 10, 50, 100, 1000, 2000, 5000, 10\,000\}$. These two sets of spectra are not distinguishable, as there are no differences in the data distributions
S2	Synthetic spectra: FWHM difference	Classify two classes of spectra each composed of one Lorentzian (both centres chosen accordingly to a normal distribution $\mathcal{N}(50, 10^2)$) and two different FWHM $\xi_1 = 7$ and $\xi_2 = 9$. We generated $N = 500$ spectra for each class. We choose $n = 5, 10, 50, 100, 1000, 2000, 5000, 10\,000$. We study accuracy of classification for various value of n
S3	Synthetic spectra with additive noise offset	Classify two classes of spectra each composed of one Lorentzian (both centres chosen accordingly to a normal distribution $\mathcal{N}(50, 10^2)$). We have chosen $n = \{5, 10, 50, 100, 1000, 2000, 5000, 10\,000\}$ and a FWHM $\xi = 7$. We generated $N = 500$ spectra for each class. We then added i.i.d. Gaussian noise with a tiny class-specific mean offset (0 *vs.* 0.01) and the same standard deviation of 0.01 (namely the noise was chosen from the two distributions $\mathcal{N}(0, 0.01^2)$ and $\mathcal{N}(0.01, 0.01^2)$)
Ra1/ Rb1	Global pixel permutation	A single, consistent random shuffle is applied to all pixels across the entire dataset. This destroys physical contiguity (peaks and baselines) while preserving the global covariance structure. This tests if the model relies on spectroscopic shapes or high-dimensional statistical geometry
Ra2/ Rb2	Independent row permutation	Every spectrum is shuffled using a unique random seed, destroying both physical contiguity and inter-pixel covariance. This serves as a control to demonstrate that model success vanishes when the high-dimensional statistical structure is eliminated.
Ra3/ Rb3	Pixel count sweep	Classify EVOO *vs.* LOO (Ra1) and EVOO *vs.* VOO (Rb1) using an increasingly high number of randomly chosen pixels $k \in [2, 35]$ from the first 50 pixels (Region ρ_1, noise only). For each k, 20 independent random subsets were tested to evaluate the climb in accuracy within chemically empty regions
Ra4/ Rb4	Feature importance: window sweep	Classify oils using non-overlapping moving windows of increasing widths $W \in 20, 50, 200, 400$ across the entire detector. This experiment tests if near-perfect separability persists in regions lacking physical signals (0–400 px) compared to peak regions (600–800 px)
Ra5/ Rb5	Feature importance: SHAP	Generate mean absolute SHAP attribution maps for both experiments (A and B) across different window sizes. This experiment identifies whether the model's "important" features correlate with chemical peaks or are distributed across the high-dimensional noise floor.

equivalent to saying that each intensity in the spectra (at each wavelength) is completely independent of the others. This is clearly not true. Although the Feldman–Hájek theorem is valid for a generic covariance (under certain assumptions, discussed in the appendices), it is an interesting question to ask what the effect of correlation between intensities at different wavelengths on the effect of high-dimensionality is. To study this, we performed the same experiments also with a Toeplitz geometric covariance. The latter is a matrix whose entries depend only on the absolute difference between their indices. This type of covariance is modelled with a parameter $\rho \in (-1, 1)$. For an n-vector $x = (x_1 \ldots x_n)$ we write $\sum(\rho, \sigma^2) \in R^{n\times n}$ with entries

$$\Sigma_{ij} = \mathrm{cov}(x_i, x_j) = \sigma^2 \rho^{|i-j|}, \quad 1 \le i,j \le n \tag{1}$$

which is Toeplitz, since each diagonal is constant. This type of covariance models situations where the correlation between intensities decreases as the difference between their wavelengths increases. Although this is an approximation in spectroscopy, where there might be large covariances between wavelength bands, it is an approximation that gives an indication on the effect of correlation and that is easily tractable mathematically and numerically.

The details of the experiments with the complete ranges of parameters tested are contained in Table 1.

3.2 Skewed normal noise classification

As we discussed in section 2.2, even when data are not normally distributed (as spectra are not), the high-dimensionality effect is still very powerful. To show this, we performed experiments using the skewed normal distribution (N4 in Table 1). This choice was guided by the analysis of noise present in the real data discussed in section 4.4.

For this experiment, we generated two classes of random noise arrays of dimensions n, analogously to the method described in the previous section, from the skewed normal distributions (SND) defined by Azzalini and Dalla Valle,[19] described below.

Let us give the mathematical form of the SND used. In its univariate version the probability density function (PDF) of a SND is given by

$$f(x) = 2\varphi(x)\Phi(\alpha x) \tag{2}$$

with

$$\varphi(x) = \frac{1}{2\pi} \mathrm{e}^{-x^2/2} \tag{3}$$

the standard normal PDF and $\Phi(x)$ the cdf of the univariate standard normal distribution. For this work, we used the multivariate version of this distribution.[19,20]

The multivariate skewed normal distribution (MSN) introduced by Azzalini and Dalla-Valle[19] is defined in its general form as follows. A random vector $X \in R^{\mathrm{p}}$ has an SND distribution with location parameter $\xi \in R^{\mathrm{p}}$, symmetric positive definite scale parameter $\Omega \in R^{\mathrm{p}\times\mathrm{p}}$, and skewness parameter $\boldsymbol{\alpha} \in R^{\mathrm{p}}$, if its multivariate PDF is

$$f_x(x) = 2\phi_{\mathrm{p}}(x; \boldsymbol{\mu}, \boldsymbol{\Omega})\Phi\{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\omega}^{-1}(x-\boldsymbol{\mu})\}, x \in R^{\mathrm{p}} \tag{2}$$

where $\varphi_{\mathrm{p}}(\cdot; \boldsymbol{\mu}, \Sigma)$ is the multivariate PDF of a p-dimensional normal distribution with mean $\boldsymbol{\mu} \in R^{\mathrm{p}}$ and $\boldsymbol{\omega} = \mathrm{diag}(\boldsymbol{\Omega})^{1/2}$. For completeness, we can write

$$\varphi_{\mathrm{p}}(x; \boldsymbol{\Omega}) = (2\pi)^{-p/2}|\boldsymbol{\Omega}|^{-1/2} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(x-\boldsymbol{\mu})\right). \tag{4}$$

Note that the matrix $\boldsymbol{\Omega}$ is a dispersion matrix and equals the covariance of X only for $\boldsymbol{\alpha} = 0$.

For our tests, we used $\boldsymbol{\mu} = \mu 1_{\mathrm{p}}$ and $\boldsymbol{\gamma} = \gamma 1_{\mathrm{p}}$ and $\boldsymbol{\Omega} = \sigma^2 \mathrm{I}_{\mathrm{p}}$. We then evaluated the separability of the classes with synthetic experiments. For each model, we generated two classes of $N = 100$ samples in $n = 50$ dimensions with coordinates i.i.d. from skew-normal distributions: a fixed base class ($\mu = 10$, $\sigma = 1$, $\gamma_1 = 0.5$) and a perturbed class obtained by shifting parameters by $\Delta\mu$, $\Delta\sigma$, and $\Delta\gamma_1$ over uniform grids. We performed three two-dimensional sweeps, ($\Delta\mu$, $\Delta\sigma$), ($\Delta\mu$, $\Delta\gamma_1$), and ($\Delta\sigma$, $\Delta\gamma_1$), and, at each grid point, trained a specific model and measured out-of-sample accuracy *via* 5-fold cross-validation. The mean cross-validated accuracies define accuracy surfaces that quantify how separability changes with differences in mean, variance, and skewness. For reproducibility, the two classes were generated with independent random seeds, and we also report results on the relative axes ($\Delta\mu/\mu$, $\Delta\sigma/\sigma$, $\Delta\gamma_1/\gamma_1$).

3.3 Synthetic spectra classification

To highlight the importance of high-dimensionality, we performed experiments with simulated spectra (S1, S2, and S3 in Table 1). In a first experiment S1 we want to show how if spectra are truly undistinguishable, then no matter the dimension or the model, classification will be no better than chance.

To show this, we simulated two classes of one-peak spectra on a discrete axis $x = (x_1, \ldots, x_n)$ (representing detector pixels, wavenumbers or wavelengths depending on the spectroscopy type). Each spectrum is a Lorentzian profile with a randomly jittered centre and a fixed full width at half maximum (FWHM), equal for the two classes. For class $k \in 1, 2$ we draw, independently for every spectrum j, a peak centre $c_j \sim \mathcal{N}(\mu, \sigma^2)$ (with values $\mu = 50$, $\sigma = 10$ and FWHM $= \xi = 7$ for numerical simulations). We used the unit-height Lorentzian.

$$L(x; c, \xi) = \frac{(\xi/2)^2}{(x-c)^2 + (\xi/2)^2},$$

We generate $N = 1000$ spectra per class with $n = 100$.

In a second experiment S2, we then studied the classification of the spectra constituted by a single Lorentzian peak with different FWHM. For a given number of dimensions $n \in \{5, 10, 50, 100, 1000, 2000, 5000, 10\,000\}$ we generate $N = 500$ spectra per class. Each spectrum is a Lorentzian profile with a centre randomly chosen from a normal distribution with parameters (μ, σ_c) = (50, 10) and a different FWHM $\xi_1 = 7$, and $\xi_2 = 9$ for the two classes. For the spectroscopist, in Fig. 2, 10

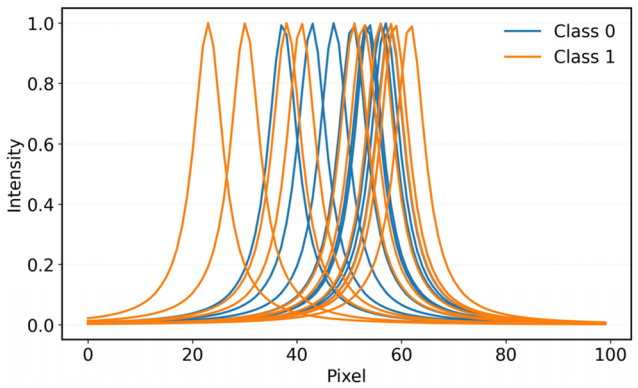

Fig. 2 Ten representative synthetic one-peak spectra per class used in the synthetic-spectra experiments. Each curve is a Lorentzian profile sampled on an n = 100-point axis, with peak centre jittered as $c \sim \mathcal{N}(50, 10^2)$. Class 0 (blue) and Class 1 (orange) differ only through the FWHM ξ_1 = 7 *vs.* ξ_2 = 9, illustrating that the two classes are visually difficult to distinguish despite being statistically separable in high dimension.

examples of each class are plotted together, to show how visually it is impossible to distinguish the two classes.

In the third experiment S3 we studied the effect of dimensionality on classification in presence of noise. In fact, in spectroscopy, for a multitude of reasons (detector dark current, electronics, *etc.*), noise is always present. We know from the experiments described here that it is possible to perfectly classify pure noise. It is an interesting question what happens if noise is added, for example, to a set of undistinguishable spectra (experiment S1).

For this experiment, we added to the spectra of experiment S1, an independent Gaussian noise that differs only in its mean between classes: for class 0 we add $\epsilon_{j,i}^{(0)} \sim \mathcal{N}(0, 0.01^2)$, and for class 1 $\epsilon_{j,i}^{(1)} \sim \mathcal{N}(0.01, 0.01^2)$. We consider N = 500 spectra per class and $n \in$ 5, 10, 50, 100, 1000, 2000, 5000.

For each n we create a balanced dataset and benchmark four standard classifiers: logistic regression (max_iter = 3000), k-nearest neighbours, a decision tree (max_depth = 5), and a random forest (100 trees, fixed seed). Performance is estimated using a 5-fold stratified cross-validation. We report the mean and standard deviation of validation accuracy across folds. All random draws use fixed pseudorandom seeds to ensure reproducibility. Because peak shape and centre statistics are identical across classes, the only class-dependent signal arises from a tiny global shift in the additive noise mean. This setting reflects common practice where models ingest spectra without explicit peak annotations; the experiment investigates how classifiers exploit minute distributional offsets when presented with high-dimensional spectral vectors.

3.4 Real spectroscopic data

The dataset used in this study consists of fluorescence spectra of Spanish olive oil acquired with a miniaturised and low-cost fluorescence-based instrument by the authors.[21] The dataset consists of 24 olive oil samples from the 2019–2020 harvest. The samples were classified by the producer, Conde de Benalúa (Granada, Spain) into 12 extra virgin oils (EVOO), 8 virgin oils (VOO), and 7 lampante oils (LOO), based on standard chemical and sensory parameters according to EU regulations. Each olive oil sample was measured undiluted and under ambient conditions. We refer the reader to the original article[21] for more details. Note that the spectra used in this work have not been normalised and have been used raw. Fig. 3 shows the spectra for the three available classes: EVOO, VOO, and LOO. In the raw data the Rayleigh scattering peak due to the excitation LED is clearly visible for EVOO and VOO but almost inexistent for LOO due to an increase absorbance in the blue-ultraviolet. Its presence, therefore, could allow models to easily distinguish, for example, EVOO from LOO. Since this work focuses on the effect of high dimensionality and minimal statistical differences on model classification performance, a spectral region between 380 and 420 nm around the excitation LED wavelength (395 nm) was eliminated. Furthermore, two classification experiments of different complexity were performed: EVOO *vs.* LOO, with clearly different spectral characteristics, and the less easily distinguishable EVOO *vs.* VOO. All experiments of the first type are indicated with Ra1, Ra2, *etc.*, while those of the second type Rb1, Rb2, *etc.* The experiments performed with these data are summarised in Table 1.

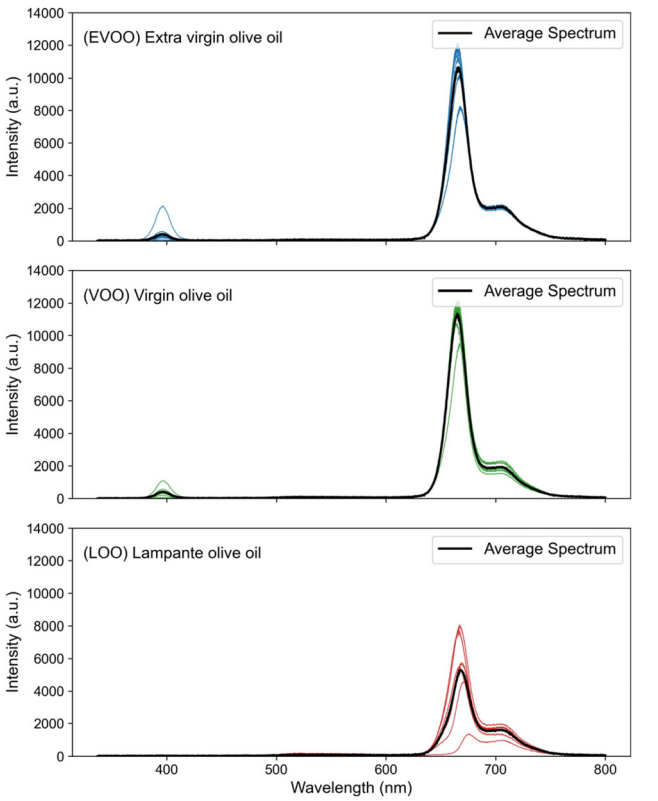

Fig. 3 Fluorescence spectra of Spanish olive oil samples classified as Extra Virgin (EVOO), Virgin (VOO), and Lampante (LOO). The region 380–420 nm indicates the Rayleigh scattering peak from the excitation LED. The black line indicates the average spectrum for each class.

Table 2 Main findings and references to results

ID	Experiment	Key finding	Reference to results
N1	Gaussian noise: $\Delta\sigma$ sweep (QDA)	Accuracy climbs monotonically with $\Delta\sigma$ and with dimension n; it reaches almost 1 already with modest gaps for n high enough. In high n white noise is easily and perfectly classifiable. We have tested $n \in \{5, 10, 50, 500\}$ and $\Delta\sigma$ from 0 to 2 for an homogoenous covariance and for a Toeplitz one with $\rho = 0.95$	Fig. 4; Methods in section 3.1; Results in section 4.1
N2	Gaussian noise: Bayes "oracle" boundary	White noise from the two distributions is almost perfectly classifiable in high enough n; accuracy goes to 1 quickly as $\Delta\sigma$ grows. We have tested $n \in \{30, 100, 500, 1000, 5000\}$ and we have varied $\Delta\sigma$ from 0 to 1.0	Fig. 5; Methods in section 3.1; Results in section 4.1
N3	Gaussian noise: accuracy *vs.* dimension n	Even small $\Delta\sigma$ becomes highly separable as n increases (dimensionality amplifies tiny distributional gaps). We have tested n from 0 to 5000 (but visualised only until 100, to make the more steeper growing curves more visible) and tested $\Delta\sigma \in \{0.1, 0.3, 0.6, 0.9, 1.2, 1.5, 2.0\}$	Fig. 6; Methods in section 3.1; Results in section 4.1
N4	Skew–normal noise: 2D parameter sweeps	Tiny shifts in mean, variance or skewness give near-perfect accuracy for most models; random forest saturates the fastest	Fig. 7; Methods in section 3.2; Results in section 4.2; Results in section 4.3
S1	Synthetic spectra: truly identical classes	No classifier exceeds chance level accuracy (0.5) confirming that without distributional differences, classification is impossible	Table 3; examples in Fig.
S2	Synthetic spectra: width difference	Validation accuracy rises as expected with n; linear and ensemble models approach 1.0 for large n similarity in the spectra	Fig. 8; Results in section 4.3
S3	Synthetic spectra with additive noise offset	High n allows a minute noise distributional difference to enable near-perfect separability; random forest reaches approximately 1.0 with very small values of n	Fig. 9; Results in section 4.3
Ra1/ Rb1	Global pixel permutation	Accuracy remains high (~82%) despite the total destruction of spectral shapes. This empirically proves the model relies on global covariance structures rather than physical spectroscopic peaks	Results in Sec. 4.4
Ra2/ Rb2	Independent row permutation	Model performance collapses to the majority-class baseline. This confirms that success in Ra3/Rb3 was due to high-dimensional statistical structure, which is destroyed by independent shuffling	Results in Sec. 4.4
Ra3/ Rb3	Pixel count sweep	Accuracy reaches > 85% using only 15–20 randomly selected pixels from the noise region (ρ_1). This confirms that non-contiguous, chemically empty data provide sufficient statistical separation in high dimensions	Fig. 10; Results in Sec. 4.4
Ra4/ Rb4	Feature importance: window sweep	High classification accuracy (~80%) is maintained across all windows, including those in the signal-free region (0–400 px). Larger windows ($W = 400$) create an accuracy plateau independent of spectral features	Fig. 12; Results in Sec. 4.4
Ra5/ Rb5	Feature importance: SHAP	SHAP attribution is distributed across the entire spectrum, often assigning higher "importance" to noise regions than to chemical peaks. This highlights the "interpretability paradox" in high dimensions	Fig. 13; Results in Sec. 4.4

4 Results

This section presents the results of the experiments. For clarity, the overview of the results for each experiment can be found in Table 2.

4.1 Gaussian noise classification

The first experiment (N1) consisted of classifying two classes of n-dimensional noise arrays from isotropic and for Toeplitz covariances: $\mathcal{N}(\mu_1, {\sigma_1}^2 I_n)$ and $\mathcal{N}(\mu_2, {\sigma_2}^2 I_n)$ and for $\mathcal{N}(\mu_1, \Sigma_1(\rho))$ and $\mathcal{N}(\mu_2, \Sigma_1(\rho))$, respectively. The results of the quadratic discriminant analysis (QDA) with a regularisation parameter equal to 0.4 are shown in Fig. 4. In this simulation, we fix $\mu_1 = \mu_2$ and sweep the standard-deviation gap $\Delta\sigma$: $= |\sigma_2 - \sigma_1|$ over [0, 2], while varying the dimensionality n (points per array). For each $\Delta\sigma$ we formed an 80/20 train-test split and reported the test accuracy. The latter is close to chance ($\approx$0.5) when $\Delta\sigma$ is close to zero and increases with both $\Delta\sigma$ and n; higher dimensions reach approximately 1.0 rather quickly with much smaller gaps (*e.g.*, hundreds of points per array achieve near-perfect accuracy for modest $\Delta\sigma$), whereas very small n require larger gaps to exceed 0.9. Panel (A) in Fig. 4 shows the results for a Toeplitz covariance with $\rho = 0.9$ and panel (B) for a homogeneous one. Notably, when considering Toeplitz matrices, the presence of correlation slows down the effect of high dimensionality (effectively it takes large gaps $\Delta\sigma$ for the same n value to reach the same accuracy value), but it does not stop it (as is expected, since the Feldman–Hájek theorem is valid for generic covariances).

In the second experiment (N2) for each variance gap value $\Delta\sigma$, we generated two classes of $N = 1000$ arrays for various values of n from isotropic Gaussians with equal mean μ and standard deviations σ_1 and $\sigma_2 = \sigma_1 + \Delta\sigma$. We evaluated the LDA decision boundary T for this setting

$$T = n \log\left(\frac{{\sigma_2}^2}{{\sigma_1}^2}\right) \frac{1}{\frac{1}{{\sigma_2}^2} - \frac{1}{{\sigma_1}^2}} \tag{5}$$

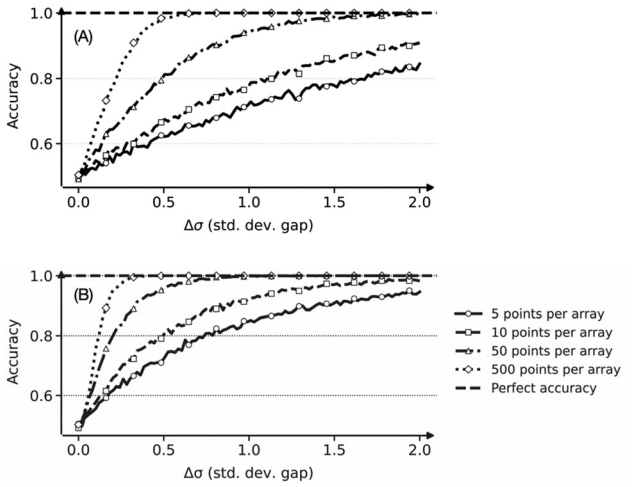

Fig. 4 Results from experiment N1. Classification accuracy of QDA (regularisation parameter equal to 0.4) as a function of the standard-deviation gap $\Delta\sigma$ between two white-noise classes with equal mean μ = 1 and baseline σ_1 = 1. Each curve corresponds to a different number of points per array ($n \in$ [5, 10, 50, 500]); at each $\Delta\sigma$, N arrays per class are generated and split 80/20 into train/test. The dashed line at 1.0 marks perfect accuracy. The results are for the test dataset. Panel (A) has been obtained with a Toeptlitz covariance with ρ = 0.95, while panel (B) with a homogeneous covariance. It is an interesting observation that adding correlations between neighbouring values, slow down the growth to perfect accuracy, but it does not stop it altogether.

which classifies a sample by thresholding its squared norm around this T value. For each $\Delta\sigma$ we formed an 80/20 train-test split and computed the test accuracy on this Bayes decision boundary. The results can be seen in Fig. 5. When $\Delta\sigma = 0$ ($\sigma_1 = \sigma_2$) the classes are indistinguishable and the accuracy is approximately 0.5; as $\Delta\sigma$ grows, the distributions become increasingly separable and the accuracy approaches1, matching the theoretical behaviour for isotropic Gaussians. Note that if we consider Toeplitz covariances, the same *slowing down* effect described for Fig. 4 appears. We have not reported these additional results to keep the length of this article reasonable.

In the third experiment (N3) we generated two classes of N = 1000 arrays from isotropic Gaussians with equal mean

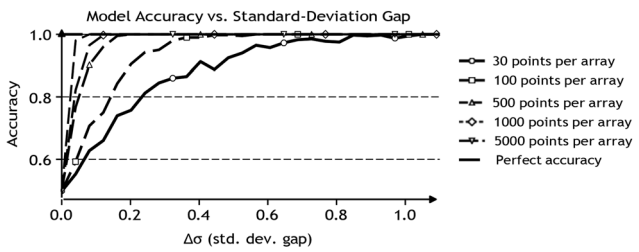

Fig. 5 Results from experiment N2. Accuracy of the Bayes classifier for two Gaussian white-noise classes with common mean μ = 10 and variances $\sigma_1^2 I_n$ *vs.* $\sigma_2^2 I_n$. The decision uses the sufficient statistic $S = \sum_{j=1}^{n} (x_j - \mu)^2$ with the LDA threshold T.

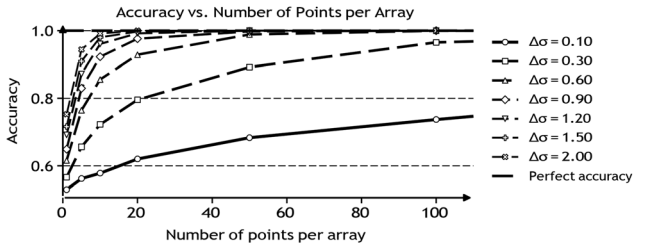

Fig. 6 Results from experiment N3. Test accuracy of QDA (reg_param = 0.4) *versus* the number of points per array n for two Gaussian white-noise classes with common mean μ and variances $\sigma_1^2 I_n$ *vs.* $\sigma_2^2 I_n$. Each curve corresponds to a different standard-deviation gap $\Delta\sigma = \sigma_2 - \sigma_1$. Datasets contain N arrays per class and are split 80/20 into train/test. The dashed horizontal line at 1.0 indicates perfect accuracy. Accuracy increases with both the number of points per array n and the variance gap $\Delta\sigma$; even small $\Delta\sigma$ yields near-perfect accuracy as n grows—showcasing how high dimensionality amplifies differences in distributions.

μ = 0 and standard deviations σ_1 = 1 and $\sigma_2 = \sigma_1 + \Delta\sigma$. The results are shown in Fig. 6, for each standard deviation gap $\Delta\sigma \in$ 0.1, 0.3, 0.6, 0.9, 1.2, 1.5, 2.0 and dimension $n \in$ 1, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000. We used QDA and calculated the test accuracy on a 20% hold-out split for each (n, $\Delta\sigma$). The curves show that accuracy increases monotonically with n and with the gap $\Delta\sigma$: larger gaps reach near-perfect accuracy at much smaller n, while very small gaps require higher dimensions to move far above chance.

4.2 Skew normal noise classification

The results of experiment N4 are shown in Fig. 7, where the cross-validated accuracy for four models trained to distinguish two classes drawn from a skew-normal law in dimension n = 50. Class 0 is fixed at (μ_1, σ_1, γ_1) = (10, 1, 0.5) and Class 1 at (μ_2, σ_2, γ_2) = ($\mu_1 + \Delta\mu$, $\sigma_1 + \Delta\sigma$, $\gamma_1 + \Delta\gamma$). Each plot sweeps two parameters while the third remains at its baseline: (A) $\Delta\mu/\mu_1$ *vs.* $\Delta\sigma/\sigma_1$, (B) $\Delta\mu/\mu_1$ *vs.* $\Delta\gamma/\gamma_1$, (C) $\Delta\sigma/\sigma_1$ *vs.* $\Delta\gamma/\gamma_1$. Contour lines help identify exact accuracy values. The colour scale ranges from chance (0.5) to perfect (1.0).

Fig. 7 shows that logistic regression achieves a near-perfect performance in most parameter ranges. A small mean shift ($\Delta\mu$) already drives the accuracy to ≈1.0, and combinations of variance and skew differences (($\Delta\sigma$, $\Delta\gamma$)) also reach 100% accuracy very quickly. Random forest shows very strong and robust results across the three plots, with broad yellow (1.0) regions, and it saturates quickly at perfect accuracy. kNN requires larger separations to leave chance accuracy and shows the steepest transition bands. This is consistent with the curse of dimensionality: local neighbourhoods are less informative in n = 50, so kNN needs larger ($\Delta\mu$, $\Delta\sigma$, $\Delta\gamma$) to separate the classes. In contrast, decision tree shows intermediate performance with more granular contours. Random forest exhibits the highest accuracy even for very small differences in the parameters.

To summarise, already in n = 50 dimensions, which is low compared to typical spectrum dimensions (of the order of 10^3), tiny discrepancies in mean, spread, or skewness already

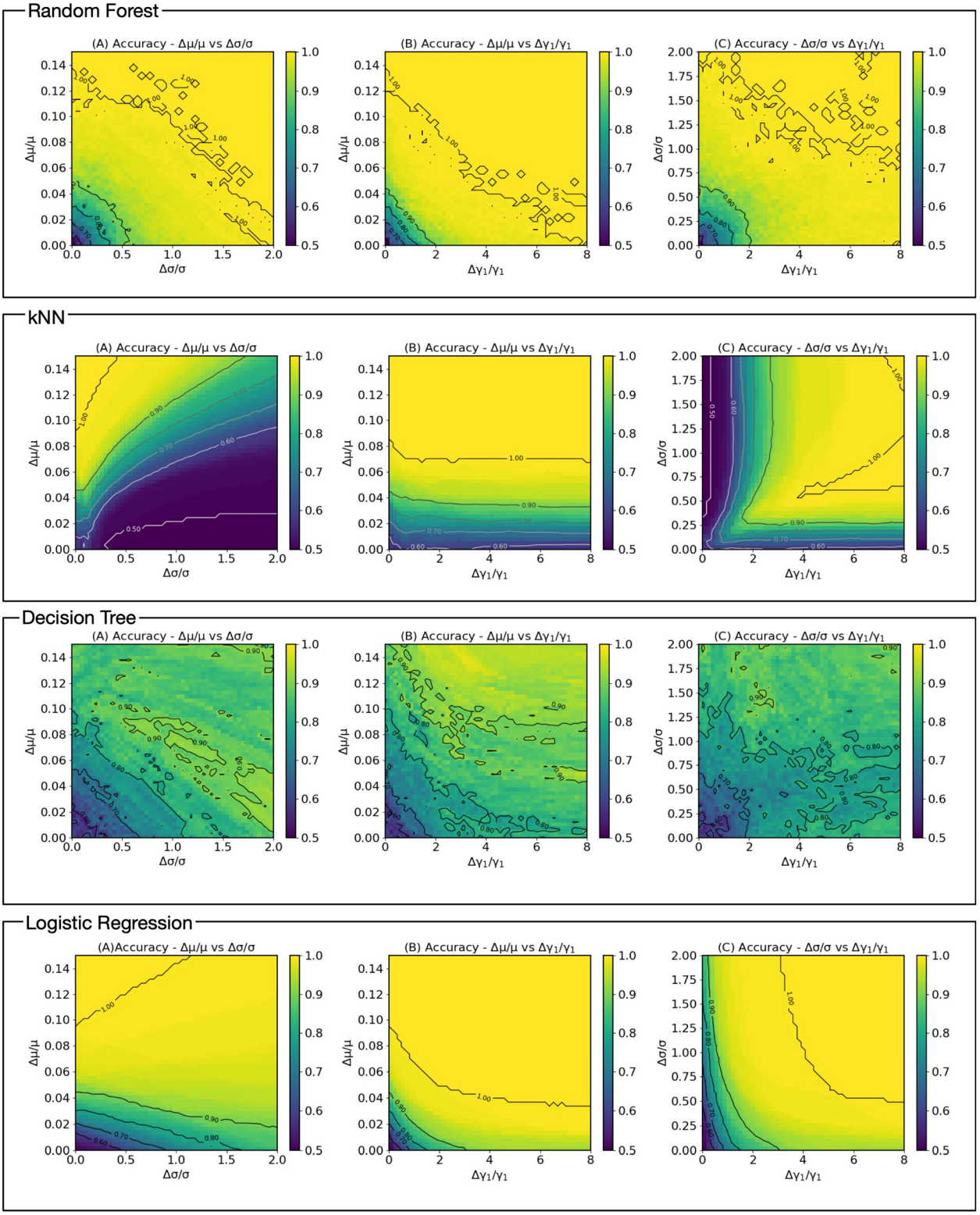

Fig. 7 Results from experiment N4. Cross-validated classification accuracy for four models (from top to bottom: Random Forest, kNN, Decision Tree, Logistic Regression) on synthetic data drawn from a skew-normal distribution in dimension $n = 50$. Each column sweeps two parameters while holding the third at its baseline: (A) $\Delta\mu/\mu_1$ *vs.* $\Delta\sigma/\sigma_1$, (B) $\Delta\mu/\mu_1$ *vs.* $\Delta\gamma/\gamma_1$, (C) $\Delta\sigma/\sigma_1$ *vs.* $\Delta\gamma/\gamma_1$, where γ is the skew (shape) parameter of the skew-normal. Class 0 is fixed at $(\mu_1, \sigma_1, \gamma_1) = (10, 1, 0.5)$ and Class 1 at $(\mu_2, \sigma_2, \gamma_2) = (\mu_1 + \Delta\mu, \sigma_1 + \Delta\sigma, \gamma_1 + \Delta\gamma)$. For every grid point we generate $N = 100$ samples per class and report the mean accuracy over 5-fold cross validation. Colour encodes accuracy (scale at right, 0.5–1.0); black contour lines mark iso-accuracy levels.

make the classes almost perfectly distinguishable. Fig. 7 shows that all four models approach the 100% accuracy in wide regions of the parameter grids.

4.3 Synthetic spectra classification

In experiment S1, as expected, no model is able to distinguish between the two classes, as the results in Table 3 show. This outcome confirms that when the underlying data distributions are truly identical, the classification task becomes statistically impossible: all models perform at chance level, with validation accuracies fluctuating around 0.5 due to random sampling. This serves as a sanity check, demonstrating that the experimental setup and models behave consistently with theoretical expectations.

In experiment S2, we observe that increasing the dimensionality enables even simple models such as logistic regression to achieve near-perfect accuracy. As shown in Fig. 8, classification performance improves steadily with the number of intensity points (or dimensions) in all models tested. These results were obtained using 5-fold stratified cross-validation and averaged over folds, with all random seeds fixed for reproducibility. These findings illustrate how, in typical spectroscopic ML analysis, where spectra are treated as high-dimensional vectors, classifiers may exploit subtle distributional differences unrelated to the actual chemical structure.

In all models, validation accuracy increases with the size of the spectrum n, reflecting the fact that the differences in FWHM become easier to detect. Linear and ensemble methods benefit the most from increasing n, with performance approaching unity for large arrays, whereas shallow trees saturate earlier. Training accuracies follow the same trend, indicating low variance for the ensemble and stable generalisation once n is sufficiently large. This controlled study isolates a single physical change (peak width) under substantial positional jitter and shows how dimensionality alone can convert a weak per-pixel signal into a highly separable representation, providing a transparent explanation for model behaviour on spectroscopic data.

The results of experiment S3 are shown in Fig. 9. The two classes differ only by a small class-specific offset in additive noise (mean 0 *vs.* 0.01 with SD 0.01); the underlying Lorentzian signal (centre distribution and width) is identical. The figure illustrates a general phenomenon in spectroscopy: when models have as input spectra as high-dimensional vectors, even subtle distributional shifts (here, a 0.01 mean offset in the noise) can become highly separable as n increases. The ensemble and nearest-neighbour methods aggregate this diffuse evidence quickly; linear models eventually catch up as the $\sqrt{n}$ gain in averaging overwhelms noise; shallow single trees remain bias-limited.

Table 3 Accuracy of multiple models for indistinguishable spectra generated as described in the text. Small deviations from 0.5 are attributable to finite-sample fluctuations

Model	Accuracy (mean ± SD)
Logistic regression	0.54 ± 0.03
K-Neighbors classifier	0.50 ± 0.04
Decision tree classifier	0.53 ± 0.04
Random forest classifier	0.51 ± 0.01

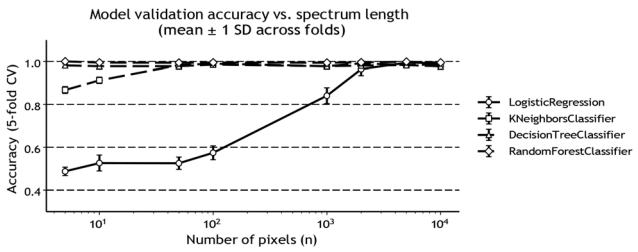

Fig. 8 Results from experiment S2. Model validation accuracy (mean ± 1 SD over 5-fold CV) *versus* spectrum length n (log scale) for four classifiers: logistic regression, k-NN, decision tree (max depth 5), and random forest (100 trees). Data consist of synthetic one-peak Lorentzian spectra: the two classes differ only in width (ξ_1 = 7 *vs.* ξ_2 = 9) while peak centres are jittered $c \sim \mathcal{N}(50, 10^2)$; 500 spectra per class are generated for each n. Markers show the mean accuracy with error bars; the legend identifies the models.

4.4 Real spectroscopic data classification

In this section, we show this phenomenon with data from real spectroscopic measurements.

4.4.1 Spectral regions and statistical differences. Let us discuss the main five different regions in the spectra.

- **Region** ρ_1: 337 nm–380 nm: this region contains only noise and no chemical fingerprint.
- **Region** ρ_2: 380 nm–420 nm: this region contains the Rayleigh scattering peak. As explained, this region has been removed from the spectra, to take away an easy way to a high accuracy for models.
- **Region** ρ_3: 420 nm–630 nm: this region contains very weak fluorescence signals, and as such chemical information, mainly due to the oxydation products in olive oil.

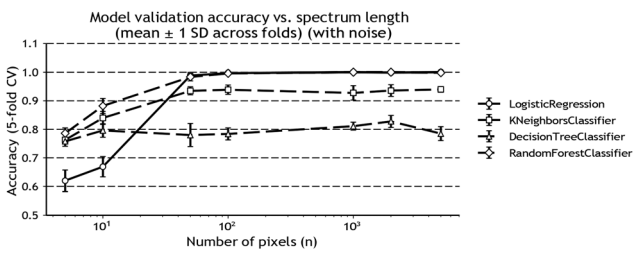

Fig. 9 Results from experiment S3. Validation accuracy (mean ± 1 SD over 5-fold CV) *versus* spectrum length n (log scale) for four classifiers on synthetic one-peak spectra with identical signal distributions but class-specific additive noise. Each class contains 500 spectra with Lorentzian FWHM ξ = 7 and centres jittered $c \sim \mathcal{N}(50, 10^2)$; noise is i.i.d. Gaussian with mean 0 (class 0) or 0.01 (class 1) and SD 0.01. Accuracy rises with n for all methods: random forests reach ≈1.0 with tens of intensity points, k-NN stabilizes around 0.93–0.95, logistic regression climbs steadily toward 1.0, while the single decision tree plateaus near 0.8.

• **Region** ρ_4: 630 nm–775 nm: this region contains the strongest fluorescence signals (due to chlorophylls) and corresponds to the strongest chemical information.

• **Region** ρ_5: 775 nm–800 nm: this region, similar to region ρ_3, contains only weak chemical information, since only the tails of the main peak are present in this region.

To study the effect of noise and dimensionality, we will use mainly regions ρ_1 and ρ_3. Region ρ_4 is less interesting for the discussion in this paper.

4.4.2 Baseline and classification accuracy with the entire spectra. First of all, it is important to establish a baseline. The easiest model we can think of is a majority class classified (classifying every sample in the majority class). Doing this will give us an accuracy of 63% for EVOO *vs.* LOO, and 60% for EVOO *vs.* VOO. Doing a classification of the spectra with a random forest model (with 100 estimators) gives us a Leave-one-out cross validation (LOO-CV) accuracy of 90% and 80%, respectively.

4.4.3 Global pixel permutation. To rigorously test whether the high accuracy is driven by the spatial arrangement of intensities (spectroscopic features) or merely by high-dimensional statistical geometry (as we claim), we performed a global pixel permutation experiment (experiment Ra1/Rb1). We applied a single, identical for all spectra random shuffle to all pixels across the entire dataset. This process preserves the statistical properties of the classes (mean and covariance) while completely destroying all physical context, such as peaks, baselines, and continuity.

Remarkably, a random forest classifier trained on these "scrambled" spectra achieved an accuracy of 82% for EVOO *vs.* LOO and 81% for EVOO *vs.* VOO. Since a peak cannot exist in a shuffled vector, this result serves as evidence: the model is not "reading" the spectra in any chemical sense. Instead, it is exploiting the concentration of measure in high-dimensional space. This confirms that in 10^3 dimensions, class-specific noise patterns and instrumental offsets become perfectly separable regardless of their physical meaning. Formulated more cautiously: the noise (intended as a non fluorescence signal) provides a higher degree of statistical separability in high-dimensional space.

This global pixel permutation experiment suggests a more cautious but powerful formulation of the high-dimensionality paradox: statistical artefacts are often 'easier' for models to exploit than chemical signals. Because instrumental noise and baseline offsets provide a consistent, high-dimensional footprint, flexible models (like random forests) can reach high accuracy by following the path of maximum statistical separation, even when the physical 'structure' of the data has been completely destroyed.

4.4.4 Independent raw permutation. To isolate whether the model relies on the global statistical structure or merely on individual sample intensities, we performed the experiment implemented an independent row permutation (experiment Ra2/Rb2). In this case, each spectrum was shuffled using a unique random seed, thereby destroying the inter-pixel covariance.

Unlike the global shuffle (which yielded ~82% accuracy), this independent shuffle caused the model performance to collapse to the baseline of the majority-class (~60%). This contrast provides a definitive multi-stage proof: the observed high accuracies in spectroscopic ML are primarily driven by the *covariance structure* of non-chemical artefacts in high-dimensional space. When this structure is preserved (global shuffle), the model succeeds without chemistry; when it is destroyed (independent shuffle), the model's "infinite-dimensional" advantage vanishes. This confirms that the Feldman–Hájek effect, governed by class-specific covariance differences, is the functional root cause of the reported high performances.

It is important to distinguish between the preservation of statistical correlation and the preservation of chemical information. A global permutation destroys all physical contiguity and spectral shapes (the "chemistry"), yet it leaves the underlying covariance matrix Σ intact (although reindexed). The fact that accuracy remains at 82% after a global shuffle proves that the model is performing a geometric separation based on these re-indexed statistical correlations rather than any recognisable spectroscopic features.

4.4.5 Pixel count sweep. The experiment R3a/R3b shows how noise in conjunction with high-dimensionality can allow a model to classify the samples. Using randomly selected subsets of intensity values from the spectral region ρ_1, previously identified as containing no chemical fingerprints, together with a random forest classifier (100 estimators), we evaluated model accuracy using a LOO-CV approach. To do so, we gradually increased the number of randomly selected pixels k drawn from the first 50 pixels, from 2 to 35 in both the EVOO *vs.* LOO and the EVOO *vs.* VOO classification. For each value of k, 20 independent random subsets were chosen.

The results, shown in Fig. 10, provide evidence for our thesis. As the number of randomly selected pixels increases, the classification accuracy climbs steeply, reaching approximately very high values already with *ca.* 15–20 pixels. This occurs despite the fact that:

1. The pixels are chosen from a region lacking any known physico-chemical signals.

2. The pixels in each subset are not necessarily contiguous, destroying any potential "hidden" spectral shapes or features.

3. The LOO-CV validation ensures that the model is generalising to unseen oil samples, rather than overfitting specific measurements.

These findings suggest that the high accuracy often reported in the application of ML in spectroscopic is not necessarily a result of the model identifying complex chemical patterns. Instead, it highlights a fundamental property of high-dimensional spaces: as the dimensionality n (here represented by k) increases, even infinitesimal distributional differences in noise or instrumental offsets between classes become almost surely separable. This experiment reinforces our cautious formulation that non-chemical noise is statistically "easier" for models to exploit than the subtle chemical signatures sought by researchers.

Fig. 10 clearly shows how a flexible model (such as a random forest) is able to use statistical differences in the data to achieve a very good accuracy, even when it should not, from

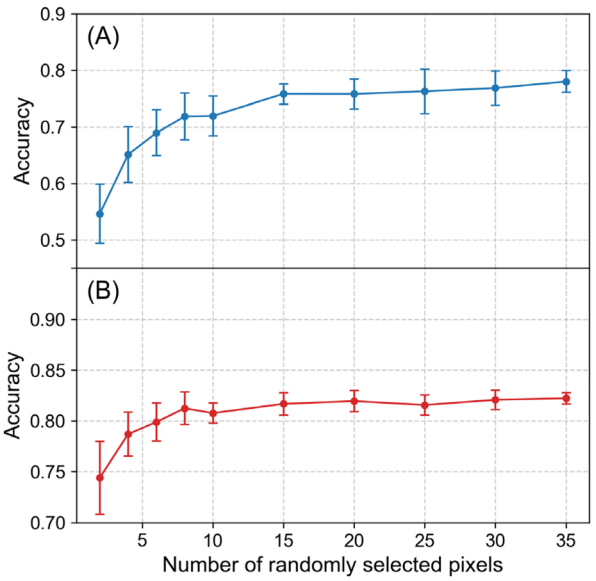

Fig. 10 Results for experiment Ra3/Rb3. LOO-CV classification accuracy as a function of the number of randomly selected pixels (k) from the spectral noise region (pixels 0–50). Panel (A) shows the results for EVOO *vs.* LOO, and Panel (B) for EVOO *vs.* VOO. Each data point represents the mean accuracy over 20 independent random subsets, with error bars indicating the standard deviation. The rapid climb to accuracies above 80%–90% using only a handful of non-contiguous, chemically empty pixels serves as evidence: the model success is driven by high-dimensional statistical separability of instrumental artefacts rather than spectroscopic chemical features.

a chemical point of view, be able to. This is naturally due to the fact that the different classes have different covariance matrices, as can be seen, for example, for EVOO and LOO in Fig. 11. Note that the large red areas in the lower right part of the covariances are related to the main peak, while the light red regions are due to stray light from the excitation LED. Note that for all the tests, we removed the Rayleigh peak from the spectra, but we wanted to include it into the covariances matrices for completeness.

4.4.6 Feature importance: window sweep. To further demonstrate how this phenomenon, we performed the experiment Ra4/Rb4 focussing on feature-importance selection. In spectroscopy, this is often done by identifying which spectral regions yield the highest classification accuracy. Since the optimal width of such regions is typically unknown, we considered several window sizes W, specifically 20, 50, 200, and 400 pixels. For each width, we used only the spectral segment within the moving window as model input (resulting in input dimensions of 10, 50, 200, and 400 pixels, respectively) and systematically slid the window along the spectrum and evaluated the performance of a random forest classifier. The results can be seen in Fig. 12 and can be summarised as follows.

- **Universality across tasks:** comparing experiment EVOO *vs.* LOO and experiment EVOO *vs.* VOO reveals a striking commonality. Despite the differing chemical complexities of these tasks, the classification accuracy in the chemically empty region (ρ_1, pixels 0–400) remains consistently high (>80%) in both cases (Panels G and H). This suggests that the model is exploiting a universal instrumental covariance structure rather than task-specific chemical markers.
- **The dimensionality plateau:** the transition from $W = 20$ to $W = 400$ illustrates the impact of dimensionality on separability. In the 400-pixel windows, the accuracy reaches a stable plateau that is indifferent to the underlying spectral profile.

These results confirm that high importance in a ML model can often be an artefact of high-dimensional geometry. When a model can achieve 80% accuracy using only randomised noise pixels (as shown in experiment Ra1/Rb1) or empty spectral windows, the standard interpretation of model weights as "chemical signatures" becomes invalid. We conclude that in 10^3-dimensional space, the most stable discriminant is frequently the global statistical fingerprint of the background, creating a deceptive "path of least resistance" that bypasses the intended chemical analysis.

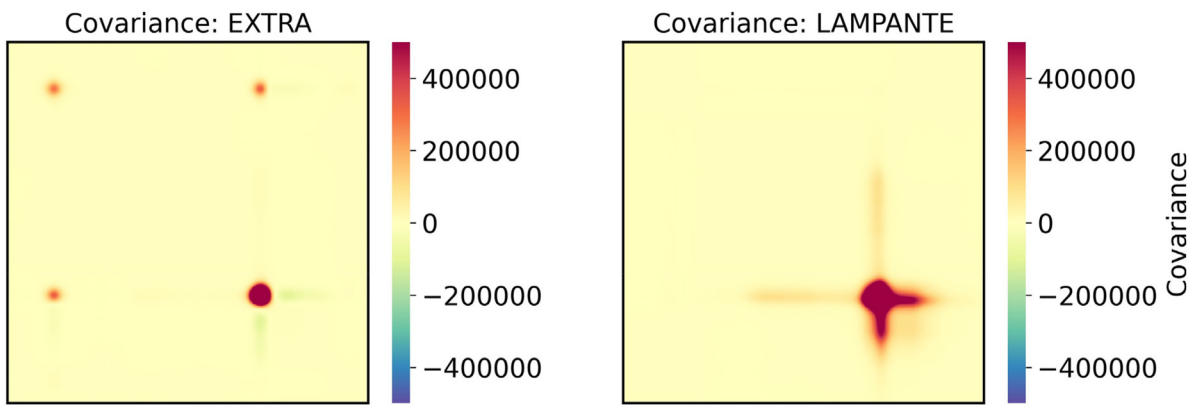

Fig. 11 Empirical covariance matrices of the fluorescence spectra for the two olive oil classes (EXTRA and LAMPANTE). Bright red areas correspond to regions of strong inter-wavelength covariances, notably around the main fluorescence peak and stray-light regions. Such covariance mismatches are sufficient, in high-dimensional space, to enable nearly perfect classification even when chemically meaningful information is absent.

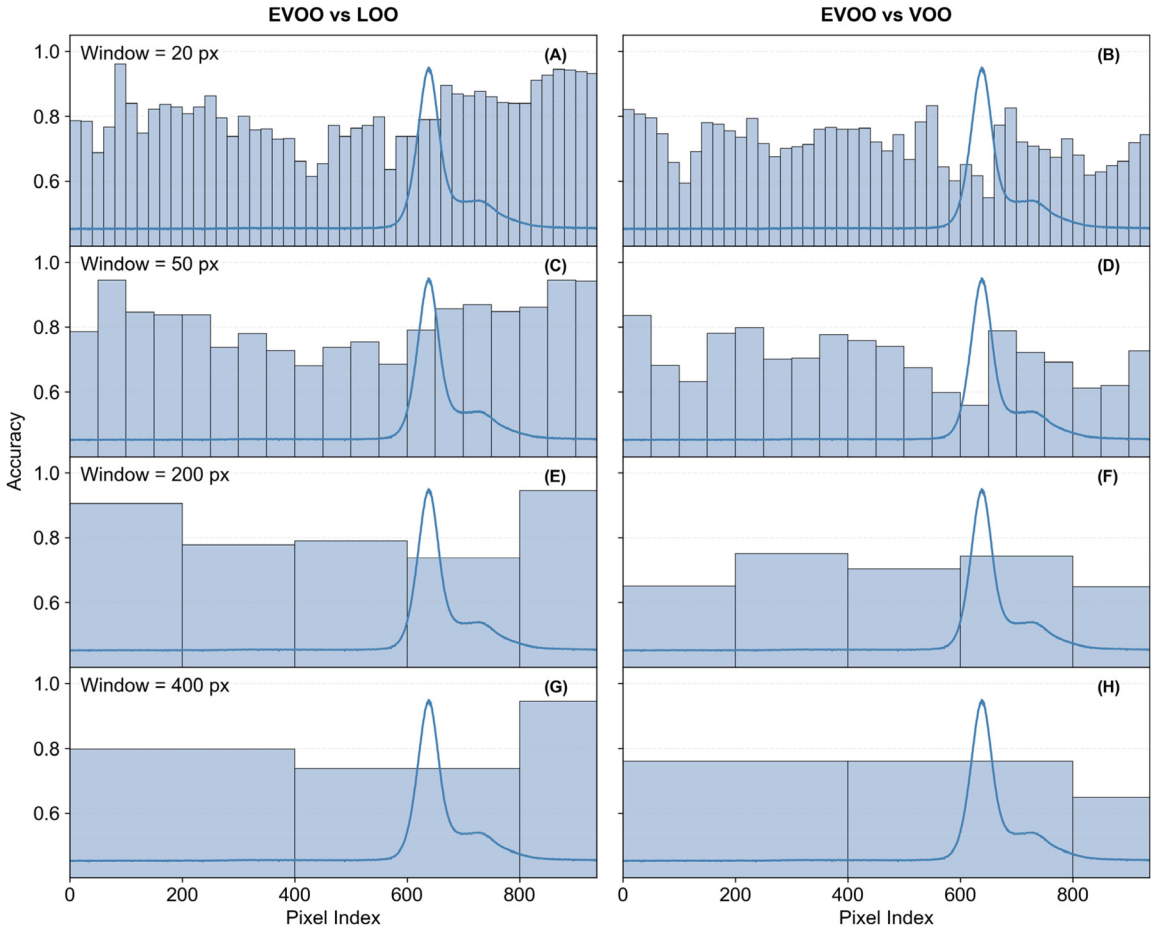

Fig. 12 Results for experiment Ra4/Rb4. LOO-CV classification accuracy is mapped across the fluorescence spectrum (gray line) using non-overlapping windows of increasing size *W*. Left Column (Panels A, C, E and G): experiment EVOO *vs.* LOO. Right Column (Panels B, D, F and H): experiment EVOO *vs.* VOO. The spectral region between 380–420 nm was explicitly removed to eliminate the Rayleigh scattering peak as a trivial discriminant. Note that even at small window sizes (20–50 px), the models achieve ~70–80% accuracy in the chemically empty region (pixels 0–400). As dimensionality increases to $W = 400$ (Panels G and H), a stable accuracy plateau of ~80% emerges across the entire detector, confirming that high-dimensional noise distributions provide sufficient statistical information for class separation, independent of physical spectroscopic features.

4.5 Feature importance: SHAP

A common approach in ML applied to spectroscopy is the use of additive feature attribution methods, such as SHAP (SHapley Additive Explanations),[22] to validate model decisions. With the experiment Ra5/Rb5 we show that SHAP is subject to the same high-dimensional "path of least resistance" as the underlying classifier. Mathematically, SHAP values are designed to distribute the total payout (the model's prediction) among the features (pixels). When classes are mutually singular due to infinitesimal shifts in the high-dimensional noise, the noise pixels collectively hold all the discriminatory power. Consequently, SHAP correctly identifies these pixels as the primary drivers of the success of the model.

However, this identifies a statistical shortcut rather than a chemical signature. In high-dimensional spaces, shortcuts are more robust and easier for the model to minimise training loss than the complex, non-linear signals of chemical peaks. Thus, a high SHAP value in a noise-dominated region is not a sign of a hidden chemical feature; it is an empirical confirmation that the model has successfully exploited the geometric separability of the instrumental background.

To explain the model's decisions, we employed SHAP. For each window size $W \in 20, 50, 200, 400$, we trained a random forest classifier on the localised spectral segment $X_{\text{start: start+W}}$. SHAP values were calculated using the TreeExplainer algorithm. To quantify the importance of a spectral window, we calculated the global mean absolute SHAP value. This metric allows us to map which spectral regions were used as primary discriminants by the model. The results can be found in Fig. 13. The feature attribution analysis was performed on both classification of EVOO *vs.* LOO and EVOO *vs.* VOO. As shown in Fig. 13, the attribution profiles reveal a significant decoupling between the spectral regions identified as important from the model and the physical chemical signal. Across all window sizes ($W = 20$ to $W = 400$ px), the model assigns high importance to spectral regions where the chemical signal is low or entirely absent. Notably, in the 400-pixel window

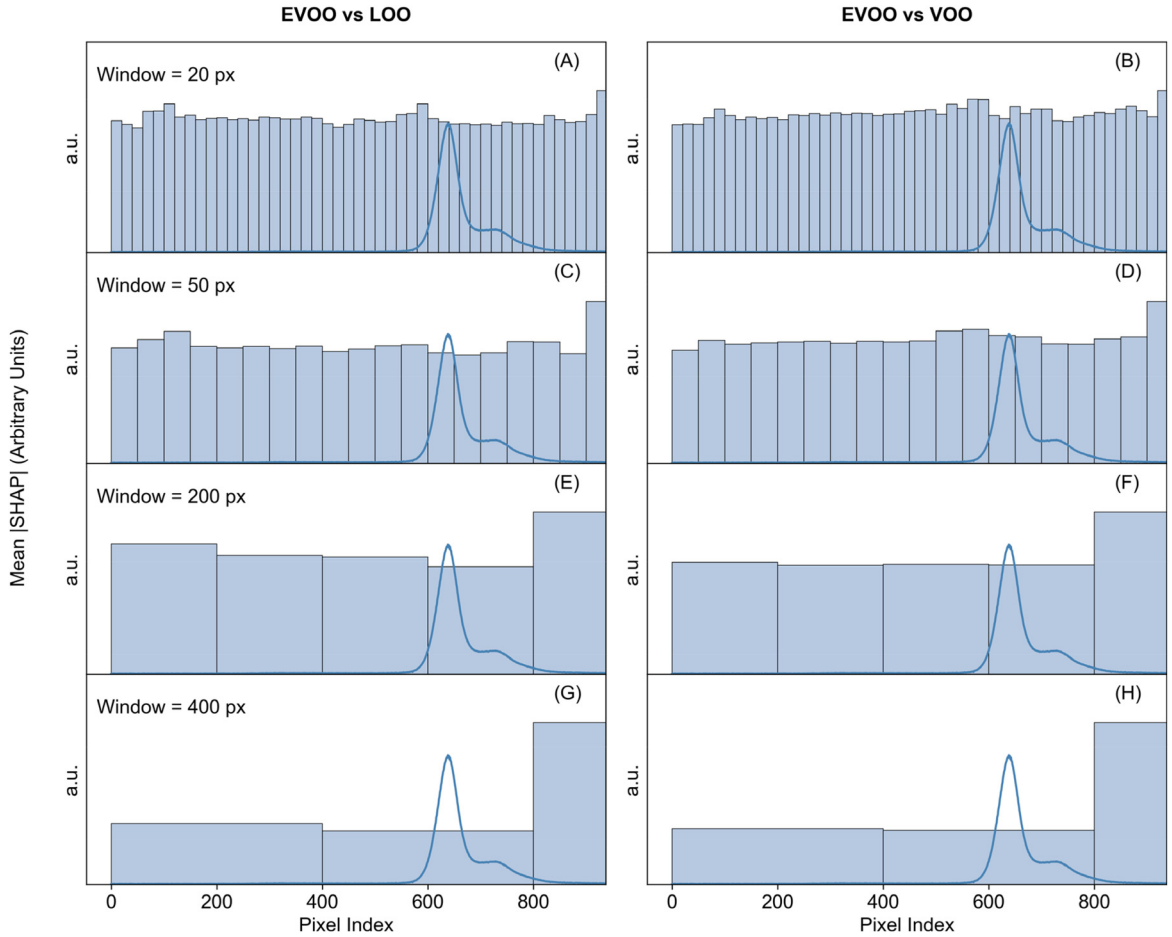

Fig. 13 Results for experiment Ra5/Rb5. Regional Feature Attribution Map. Mean absolute SHAP values are presented in arbitrary units (a.u.) to facilitate the comparison of relative feature importance across varying window sizes (*W*). Left Column (Panels A, C, E and G): experiment EVOO *vs.* LOO. Right Column (Panels B, D, F and H): experiment EVOO *vs.* VOO. The spectral region between 380–420 nm was explicitly removed to eliminate the Rayleigh scattering peak as a trivial discriminant. The persistence of high attribution in the noise-only region (pixels 0–400) across both experimental tasks indicates a reliance on high-dimensional statistical shortcuts rather than physical chemical signatures.

regime (Panels G and H), the importance assigned to the noise-only region (pixels 0–400) is comparable to, or even exceeds, the importance assigned to the primary fluorescence peaks (pixels 600–800). This indicates that the model is not relying on specific chemical markers, but is instead utilising the global high-dimensional background as a primary discriminant.

5 Discussion and practical implications for spectroscopic modeling

The results presented in this work demonstrate that the apparent success of machine learning in spectroscopy is often a consequence of the inherent high dimensionality of spectral data. In such spaces, even tiny distributional differences, such as small variations in noise or background, make the data perfectly separable as the number of spectral points grows. As a result, classifiers can achieve seemingly perfect accuracy without learning chemically meaningful features, relying instead on subtle artefacts or instrument-specific noise patterns.

When feature selection or wavelength band selection is applied to spectroscopy data, the high dimensionality of spectra can produce misleading results. Since classifiers can exploit minute distributional differences, even in spectral regions that contain no physico-chemically meaningful information, commonly used importance approaches often highlight bands that are merely correlated with noise patterns or instrument artefacts. For instance, a random forest might consistently assign high importance to regions far from characteristic peaks, not because those wavelengths encode chemical signatures but because small statistical fluctuations in those regions suffice to separate classes in high-dimensional space. This effect can lead spectroscopists to misinterpret the outcome of ML models. A feature ranking that emphasises noise-driven regions can be taken as evidence of a new "hidden" marker, while in reality the model is simply exploiting spurious differences in baseline or detector noise. As a

result, band-selection workflows risk reinforcing artefacts rather than guiding the discovery of meaningful chemical or physical features. This danger is especially acute when spectra are normalised or preprocessed, since those steps may amplify or redistribute noise in ways that make certain bands appear systematically discriminative.

Therefore, great caution is required when interpreting the output of band-importance methods. Any highlighted region should be cross-validated against established chemical knowledge or verified with independent measurements. Without this step, spectroscopists risk drawing incorrect conclusions, such as attributing predictive power to wavelength regions that carry no true spectroscopic signal. The findings of this study suggest that feature selection in spectroscopy, if performed without domain knowledge, can easily mislead and produce models that generalise poorly across instruments, conditions, or sample sets.

Dark signal and stray light must be mentioned in this context. In fact, they act as structured "noise" that can differ by instrument, session, or acquisition order and can alone enable near-perfect separation in high dimension and mislead band-importance analyses (*e.g.*, highlighting off-peak regions with no chemical content). Models trained under such conditions may fail to generalise across instruments or setups, despite excellent internal validation. Practically, this calls for rigorous controls: randomise acquisitions across classes, replicate across instruments/sessions, evaluate with leave-instrument/session-out validation, and verify that accuracy collapses when noise statistics are equalised (*e.g.*, per-scan mean/variance standardisation or explicit dark/stray-light correction). Only signals that remain discriminative under these checks should be interpreted as chemically meaningful.

Ultimately, this work should not be interpreted as a general refutation of machine learning in spectroscopy, but rather as a call for a more rigorous, evidence-based framework for model validation; we propose that high classification accuracy must be accompanied by regional sensitivity audits—such as the windowed SHAP analysis and global shuffle tests presented here—to ensure that model success is derived from verifiable chemical signatures rather than high-dimensional statistical shortcuts.

When applying machine learning to spectroscopy, it is essential to check whether models are separating classes based on chemically meaningful information or on trivial artefacts. A useful diagnostic is to test performance on wavelength regions that should be indistinguishable and contain no chemical signal; if the model still performs above chance, then separability is likely driven by noise or measurement artefacts.

Preprocessing choices also play a critical role. Steps such as baseline subtraction or normalisation can unintentionally amplify or suppress noise patterns, creating the illusion of meaningful separation. Similarly, spectral band importance methods (*e.g.* feature maps from random forests, SVMs, or SHAP values) may highlight regions that correspond to noise rather than true peaks, and therefore these results should be interpreted with great caution.

In conclusion, spectroscopists must remain aware that models trained on data from one instrument or measurement setup may not generalise to another. Retraining or re-validation is essential when changing experimental conditions. The safest approach is to combine machine learning with domain knowledge of peak positions, line shapes, and chemical constraints, and to begin with synthetic or well-characterised spectra where the discriminative features are known. This provides a baseline to ensure that models are learning physically relevant information rather than statistical quirks of the dataset.

5.1 Distinguishing overfitting from high-dimensional separability

It is essential to distinguish between classical overfitting and the phenomenon of high-dimensional separability described in this work. Although both can result in deceptively high accuracy, their underlying mechanisms differ.

- **Overfitting** typically occurs when the complexity of a model (number of parameters) is too high relative to the number of samples N. In this state, the model "memorises" specific noise fluctuations in the training set that do not exist in the population.
- **High-dimensional separability** (the Feldman–Hájek effect) is a geometric property where two distributions become mutually singular as the number of dimensions n increases. In this case, the model is not necessarily "memorising" noise; rather, it is correctly identifying that in 10^3 dimensions, the classes occupy disjoint regions of space due to minute differences in their global covariance or mean.

A key diagnostic to distinguish the two is the *rate of convergence* to perfect accuracy. In classical overfitting, accuracy usually improves as the number of samples N decreases (making the "memorisation" easier). In contrast, high-dimensional separability is driven by the number of pixels n. As shown in our experiments (Fig. 6 and 10), even with a fixed or increasing sample size, accuracy increases steadily as more spectral points are added. Furthermore, our "shuffle" experiments demonstrate that the model is exploiting global statistical distributions, which are properties of the population, rather than just local pixel-wise noise.

5.2 Generality across spectroscopic techniques

Although real-world validation in this study uses fluorescence spectra, the theoretical framework grounded in the Feldman–Hájek theorem and the concentration of measure is inherently platform-independent. The phenomenon of high-dimensional separability is a property of the data's geometry rather than its physical origin. Consequently, these findings are highly relevant to other common techniques, such as Near-Infrared (NIR) or Raman spectroscopy. In NIR spectroscopy, specifically, spectral features are characterised by broad, overlapping features rather than sharp, distinct peaks. This low "chemical contrast" makes NIR models particularly susceptible to the "path of least resistance" described in this work: a high-dimensional classifier may easily bypass the subtle chemical signals in

favor of infinitesimal but perfectly separable differences in the instrumental noise floor or baseline offsets.

5.3 Model complexity and susceptibility to dimensional effects

It is important to note that the theoretical framework presented here does not imply that all ML models are destined to fail in high-dimensional spectroscopic applications. Rather, it suggests that different architectures exhibit varying levels of susceptibility to these geometric effects. Highly flexible, non-linear models—such as deep neural networks or random forests, are particularly adept at identifying and exploiting subtle, high-dimensional covariance patterns. Because these models prioritise the minimisation of training error, they naturally gravitate toward the most statistically separable features, which, in high dimensions, are frequently instrumental artefacts or noise distributions rather than complex chemical signatures. In contrast, simpler linear models or those incorporating strong domain-specific regularisation may be less prone to "meaningless" success, provided they are constrained to look for physically plausible features.

Also it is important to note, that even in high dimensions and with clear statistical differences, it is possible that specific model classes will not reach a high accuracy. The fact that two classes are, in principle, perfectly separable it does not mean that every model can do that, or that it is an easy task. The decision bounday may be too complex for specific model classess to detect, and thus even very flexible models might have a low accuracy in classification tasks, even if in high dimensions.

6 Conclusions

This study provides a solid theoretical and empirical refutation of the assumption that high classification accuracy is a sufficient proxy for a model learning from physico-chemical information in spectroscopy. By linking the Feldman–Hájek theorem to experimental fluorescence data, we have shown that the "Path of Least Resistance" for a ML model to obtain high accuracy can be the background rather than the intended chemical signal.

We hope that our findings serve as a useful framework for the field: we propose that the standard for "model success" must be elevated from simple cross-validation accuracy to a rigorous **Regional sensitivity audit**. The windowed SHAP importance maps, the global shuffle tests, and the physical feature-removal protocols developed here provide a possible blueprint for a new generation of "physically-aware" machine learning. By adopting these stress-tests, the spectroscopy community can safeguard against the publication of non-replicable "phantom" models and ensure that the power of artificial intelligence is harnessed to uncover genuine molecular insights rather than high-dimensional geometric artefacts.

Note that while this study utilises fluorescence spectra characterised by broad features and significant instrumental background, it is important to note that the impact of high-dimensional statistical shortcuts may vary in datasets containing more dense chemical information, such as the infrared fingerprint regions of biological tissues, where the higher 'chemical contrast' might offer more robust physical discriminants.

Author contributions

Umberto Michelucci: conceptualisation, data curation, methodology, formal analysis, investigation, writing – original draft, writing – review & editing. Francesca Venturini: conceptualisation, methodology, validation, writing – original draft, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

This study was carried out using synthetic data. The data generation is explained in the paper. The olive oil data were previously described in F. Venturini, M. Sperti, U. Michelucci, I. Herzig, M. Baumgartner, J. P. Caballero, A. Jimenez and M. A. Deriu, *Foods*, 2021, **10**, 1010 and are openly available in "Dataset of Fluorescence Spectra and Chemical Parameters of Olive Oils" at https://data.mendeley.com/datasets/thkcz3h6n6/6.

Acknowledgements

This project/research has received funding from the European Union's Horizon Europe Research and Innovation Programme under the Marie Skłodowska-Curie Actions (MSCA) Staff Exchanges, Grant Agreement No. 101236434 (SMARTOLIVE).

References

1 H. Mark and J. Workman Jr, *Chemometrics in spectroscopy*, Elsevier, 2010.
2 C. L. M. Morais, K. M. G. Lima, M. Singh and F. L. Martin, *Nat. Protoc.*, 2020, **15**, 2143–2162.
3 S. Guo, J. Popp and T. Bocklitz, *Nat. Protoc.*, 2021, **16**, 5426–5459.
4 J. Workman, Jr., *Demystifying the Black Box: Making Machine Learning Models Explainable in Spectroscopy*, https://www.spectroscopyonline.com/view/demystifying-the-black-box-making-machine-learning-models-explainable-in-spectroscopy, 2025, Last acccessed on 13th Oct. 2025.
5 A. Mamalakis, E. A. Barnes and I. Ebert-Uphoff, *Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience*, *arXiv*, 2022,

preprint, arXiv:2208.09473 [physics], DOI: 10.48550/arXiv:2208.09473.
6 J. Contreras and T. Bocklitz, *Pflügers Arch.*, 2025, **477**, 603–615.
7 M. Zehtabvar, K. Taghandiki, N. Madani, D. Sardari and B. Bashiri, *Spectrosc. J.*, 2024, **2**, 123–144.
8 D. Steinmann, F. Divo, M. Kraus, A. Wüst, L. Struppek, F. Friedrich and K. Kersting, *Navigating Shortcuts, Spurious Correlations, and Confounders: From Origins via Detection to Mitigation*, *arXiv*, 2024, arXiv:2412.05152 [cs], DOI: 10.48550/arXiv:2412.05152.
9 L. Samhita and H. J. Gross, *Commun. Integr. Biol.*, 2013, **6**, e27122.
10 R. Houhou and T. Bocklitz, *Anal. Sci. Adv.*, 2021, **2**, 128–141.
11 C. T. J. Alkemade, W. Snelleman, G. D. Boutilier, B. D. Pollard, J. D. Winefordner, T. L. Chester and N. Omenetto, *Spectrochim. Acta, Part B*, 1978, **33**, 383–399.
12 J. Feldman, *Pac. J. Math.*, 1958, **8**, 699–708.
13 J. Hájek, *Czechoslovak Math. J.*, 1958, **8**, 610–618.
14 S. Axler, *Measure, Integration & Real Analysis*, Springer International Publishing, Cham, 2020, vol. 282.
15 U. Michelucci, *The Feldman-Hájek Dichotomy for Countable Gaussian Mixtures and their Asymptotic Separability in High Dimensions*, *arXiv*, 2026, arXiv.2601.03911 [math.ST], DOI: 10.48550/arXiv.2601.03911.
16 J. P. Ioannidis, *Lancet*, 2005, **365**, 454–455.
17 R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan and Y. Wang, *Nat. Rev. Cancer*, 2008, **8**, 37–49.
18 E. Vul, C. Harris, P. Winkielman and H. Pashler, *Perspect. Psychol. Sci.*, 2009, **4**, 274–290.
19 A. Azzalini and A. D. Valle, *Biometrika*, 1996, **83**, 715–726.
20 S. Mondal, R. B. Arellano-Valle and M. G. Genton, *Stat. Pap.*, 2024, **65**, 511–555.
21 F. Venturini, M. Sperti, U. Michelucci, I. Herzig, M. Baumgartner, J. P. Caballero, A. Jimenez and M. A. Deriu, *Foods*, 2021, **10**, 1010.
22 S. M. Lundberg and S.-I. Lee, *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 4765–4774.